A GENERALIZATION OF BAYESIAN ESTIMATION IN FINITE MIXTURE OF DISTRIBUTIONS

Bertha Liliana Garrido Lopera

01-196914

Thesis to obtain the degree: Doctor in Statistical Science

DIRECTOR:

EDILBERTO CEPEDA CUERVO PhD.

UNIVERSIDAD NACIONAL DE COLOMBIA

FACULTAD DE CIENCIAS

DEPARTAMENTO DE ESTADÍSTICA

Bogotá, 2010

Contents

1 HETEROSCEDASTIC EXPONENTIAL-NORMAL MIXTURE MODELS: A NEW BAYESIAN APPROACH **7**

1.1 **Introduction** 8

1.2 **Bayesian Methodology** 10

1.3 A Simulation Study 16

1.3.1 First Simulation 16

1.3.2 Comparative simulations 18

1.4 Applications 19

1.4.1 Insects 19

1.4.2 DDT 22

1.5 Conclusions and Further Results 24

2 MIXTURE OF DISTRIBUTIONS IN THE BIPARAMETRIC EXPONENTIAL FAMILY: A BAYESIAN APPROACH **33**

2.1 **Introduction** 34

2.2 **The Model** 36

2.3 **Bayesian Methodology** 39

2.3.1 Prior distributions 39

2.3.2 Conditional Posterior Distributions 41
2.3.3 Proposed Bayesian methodology 43
2.4 **Simulations** 47
2.4.1 First simulated study 48
2.4.2 Second simulated study 49
2.4.3 Third simulated study 53
2.5 Applications 58
2.5.1 Home valuations 58
2.6 Conclusions 59
2.7 Appendix: Graphs and histograms of the parameters in the third simulation and the home valuations example* 66

3 HETEROSCEDASTIC WEIBULL-NORMAL MIXTURE MODELS: A BAYESIAN APPROACH 73

3.1 **Introduction** 73
3.2 **The Model** 75
3.3 **Bayesian Methodology** 78
3.3.1 Prior distributions 78
3.3.2 Conditional posterior distributions of the parameters . 80
3.3.3 Bayesian proposed methodology 82
3.4 **Simulations** 86
3.4.1 **First simulated study** 86
3.4.2 **Second simulated study** 89
3.5 **Applications** 92
3.5.1 **Example 1** 92
3.5.2 **Example 2** 94

3.6 Conclusions 99

4 MIXTURE MODELS APPLIED TO TAR MODELS AND SPATIAL STATISTICS 105

4.1 **Introduction** 105

4.2 **Application to TAR models** 106

4.2.1 TAR models 106

4.3 **The mixture model** 107

4.3.1 **Canadian Lynx data set** 108

4.3.2 **Sunspot numbers** 111

4.4 **Application to Spatial Statistics** 116

4.4.1 **The model** 118

4.4.2 **Estimation results** 120

Agradecimientos

Agradezco al Dr. Edilberto Cepeda por su enseñanza y apoyo.

Palabras clave

Metodologa bayesiana, mixtura finita de distribuciones, familia exponencial biparamtrica, familia weibull biparamtrica.

Resumen

Se emplea metodologa bayesiana, especficamente el muestreador de Gibbs y el algoritmo de Metropolis-Hastings, para estimar los parmetros en una mixtura finita de distribuciones pertenecientes a la familia exponencial biparamtrica, o a la familia de weibull biparamtrica, modelando media y varianza de las distribuciones involucradas. En una mixtura de k distribuciones hay m de un familia de distribuciones y $k-m$ de otra familia de distribuciones, $m = 0, \ldots, k$. Las distribuciones que se trabajaron en los algoritmos fueron especficamente, normal y exponencial, normal y gama, y normal y weibull. La media y la varianza se modelaron con regresiones lineales y no lineales con un nmero arbitrario de covariables. Se aplic la metodologa bayesiana a la mixtura finita para modelar ejemplos tpicos de la estadstica espacial y de los modelos TAR de series de tiempo no lineales.

Key words

Bayesian methodology, finite mixture of distributions, biparametric exponential family, biparametric weibull family

Abstract

Bayesian methodology is employed, mainly the Gibbs sampler and the Metropolis-Hastings algorithm, to estimate the parameters in a finite mixture of distributions belonging to the exponential biparametric family, or the biparametric weibull family of distributions, modeling the mean and the variance of all the distributions involved. In a mixture consisting of k distributions, there are m from one family and $k-m$ from another family, $m = 0, \ldots, k$. The algorithms worked with distributions from the normal and exponential families, normal and gamma families, and normal and weibull families. The mean and the variance, with an arbitrary number of covariates, were modelled with linear and non linear regressions. Bayesian methodology was applied to finite mixtures to model typical examples from spatial statistics and from non linear time series TAR models.

Introduction

One can think of many data sets to be fitted by mixture models, to cite a few, some economic variables in a country where there is a big gap between rich and poor; some economic variables in a continent where there are geographical differences, for example, north and south; survival time to different treatments of an illness.

The purpose in this work is to generalize the finite mixture of distributions mainly in two ways. One way is to allow in the mixture any finite number of distributions from two biparametric subfamilies, the other way is the modelling of the mean and the variance in all the distributions involved in the mixture. Linear and non linear regression equations with any number of covariates model the mean and the variance.

Many authors have worked with mixture models and will be referenced later, but it is worth mentioning some authors like Casella, Mengersen, Robert and Diebolt, who have consistently worked with finite mixture models, and have also used bayesian techniques of estimation.

In chapter one the finite mixture of normal and exponential distributions is presented as a preamble to a more general mixture in chapter two, which is the mixture of two distributions from the biparametric exponential family, taking distributions from the normal and gamma families to ilustrate the procedure. In chapter three the mixture involves distributions from the normal and biparametric weibull families, mixture which could also have any distribution from the biparametric family of extreme value distributions.

Bayesian methodology to estimate the parameters involved is employed, mainly Gibbs sampler and Metropolis-Hastings algorithm. In the mixture of normal and weibull distributions, numerical methods have to be employed to solve for the original parameters of the weibull distributions.

In chapter four the mixture models of chapter two are employed in two important statistical subjects which are TAR models and spatial statistics, not being the objective to get deep in the theory of these subjects, but to take typical examples from those fields and model them with mixture models using bayesian methodology. A very good treatment of non linear time series models can be found in Tong's book (1995). Moreno and Vayá (2000) are the authors of a short and illustrative book in spatial econometrics.

Chapter 1

HETEROSCEDASTIC EXPONENTIAL-NORMAL MIXTURE MODELS: A NEW BAYESIAN APPROACH

Summary: In this chapter, we introduce a Bayesian analysis for mixture of distributions belonging to the exponential family. As a special case we consider a mixture of normal exponential distributions including joint modelling of the mean and variance. We also consider joint modelling of the mean and variance heterogeneity.Markov Chain Monte Carlo (MCMC) methods are used to obtain the posterior summaries of interest.We also introduce the analysis of real data sets to illustrate the proposed methodology.

Key-words:mixture models, variance heterogeneity, Bayesian methods, MCMC simulation.

1.1 Introduction

Many authors have been working with Bayesian estimation of mixture of distributions (see for example, Mengersen and Robert, 1996; Carroll, Roeder and Wasserman, 1996; Roeder and Wasserman, 1997; Titterington, Smith and Makov, 1985; Robert, 1996). In the parametric mixture model, the component distributions are from a parametric family, with unknown parameters θ_j :

$$f_X(x) = \sum_{j=1}^{k} a_j f_Y(x; \theta_j). \tag{1.1}$$

for some mixture proportions $0 \leq a_j \leq 1$ where $a_1 + \cdots + a_k = 1$.

Observe that if the number of mixture components in (4.1) is known, we have a parametric mixture model. Stephens (1997, 2000), Polymenis and Titterington (1998), Richardson and Green (1997), Diebolt and Robert (1994), Dey, Kuo and Sahu (1995), Nobile (2004), Nobile and Fearnside (2007), are recent works that address this subject.

Some of the works in the extensive literature about the subject are mentioned below. A general treatment for mixture models is given in Richardson and Green (1997) with an unknown number of distributions, all of them from the same class of normal distributions, but not considering the presence of covariates. Peng et al (1996) use mixture of normals in the presence of covariates, even in the incidence probabilities, but not modelling the variance. Achcar et al (1999a, b) mix normal and exponential distributions, with the variance of the normal distributions involved, homoscedastic. Grodzenskii and Domrachev (2002) mix an exponential and a weibull distribution considering one parameter for the exponential distribution, two for the weibull distribution and one for the incidence probability, using maximum likelihood to estimate them. Hazan et al (2003) mix two exponential power distributions with no presence of covariates, applying the Gibbs sampler to estimate the parameters. Hurn et al (2003) minimize a loss function to estimate mixtures of regressions, with homoscedastic variance in each mixture component. Dias and Wedel (2004) compare three typical estimating procedures, EM, SEM and MCMC, in Gaussian mixtures with no covariates. Grün and Leisch (2007) propose finite mixture of generalized linear regression models, using a linear regression to model the mean via a link function, with homoscedastic variance, estimating a finite mixture of normal distributions using the EM algorithm. Frühwirth-Schnatter and Pyne (2010) use bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions, with no presence of covariates, modelling Alzheimer's desease data with a skew-normal mixture.

In this chapter, we extend the Bayesian methodology proposed in Cepeda and Gamerman (2001) to the study of mixture of distributions belonging to the exponential family. Thus, for example, we consider the mixture of exponential and normal distributions, including joint modelling of the mean and the variance. We also consider joint modelling of the mean and variance heterogeneity in the case of mixture of normal-normal, gamma-normal or gamma-gamma distributions, among others.

As a special case, we consider the mixture of exponential and normal distributions. In this model we have two explanatory variables, X and W, where X is related to the means of both distributions as well as for the incidence probabilities, and W is related to the variance of the normal distribution.

Thus, if $\boldsymbol{\beta} = (\beta_0, \beta_1)'$, $\boldsymbol{\lambda} = (\lambda_0, \lambda_1)'$, $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)'$ and $\boldsymbol{\tau} = (\tau_0, \tau_1)'$, the variable of interest, Y, has a density function given by:

$$
\begin{aligned}
f(y|x, w, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\tau}) = & \frac{e^{(\tau_0+\tau_1 x)}}{1+e^{(\tau_0+\tau_1 x)}} \frac{1}{\sqrt{2\pi\sigma^2(w)}} e^{-\frac{1}{2}\left(\frac{y-(\beta_0+\beta_1 x)}{\sigma(w)}\right)^2} \\
& + \frac{1}{1+e^{\tau_0+\tau_1 x}} \mu(x) e^{-\mu(x)y} I_{(0,\infty)}(y) \qquad (1.2)
\end{aligned}
$$

where $\sigma^2(w)$ =exp$(\gamma_0+\gamma_1 w)$, $\mu(x)$=exp$(-(\lambda_0+\lambda_1 x))$. The method proposed in the bayesian methodology can be easily extended to any link function, one of them the exponential function, which guaranties the positiveness of μ and σ^2.

Posterior summaries of interest for mixture models usually have been obtained using MCMC (Markov Chain Monte Carlo) methods as the popular Gibbs sampling algorithm (see, for example, Gelfand and Smith, 1990) and the Metropolis-Hastings algorithm (see, for example, Chib and Greenberg, 1995 or Gamerman, 1997).

The chapter is organized as follows: in section 2, we propose the Bayesian methodology to fit the mixture model to the data; in section 3, we have the results from a simulation study; in section 4, we present two applications; finally, in section 5 we give some general conclusions.

1.2 Bayesian Methodology

Assuming the mixture model introduced by (1.2), the likelihood function considering a vector of observations $\mathbf{y} = (y_1, \ldots, y_n)'$ with corresponding covariates $\mathbf{x} = (x_1, \ldots, x_n)'$, $\mathbf{w} = (w_1, \ldots, w_n)'$, is given by

$$L(\mathbf{y}/\mathbf{x}, \mathbf{w}, \mathbf{\Theta}) =$$
$$\prod_{i=1}^{n} \left(\frac{e^{\tau_0+\tau_1 x_i}}{1+e^{\tau_0+\tau_1 x_i}} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}\left(\frac{y_i-(\beta_0+\beta_1 x_i)}{\sigma_i}\right)^2} + \frac{1}{1+e^{\tau_0+\tau_1 x_i}} \mu_i e^{-\mu_i y_i} I_{(0,\infty)}(y_i) \right) \tag{1.3}$$

where σ_i^2 =exp$(\gamma_0 + \gamma_1 w_i)$, μ_i=exp$(-(\lambda_0 + \lambda_1 x_i))$.

Under a Bayesian approach, we observe that the joint posterior distribution for the parameters of the model has a complex form, with some difficulties to obtain the posterior summaries of interest. A simplification is obtained by the introduction of latent variables (see for example, Tanner and Wong, 1987; or Casela et al, 2002) z_i, which are defined as Bernoulli variables as follows, let us define h_i by,

$$h_i = \frac{\frac{e^{(\tau_0+\tau_1 x_1)}}{1+e^{(\tau_0+\tau_1 x_i)}} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}\left(\frac{y_i-(\beta_0+\beta_1 x_i)}{\sigma_i}\right)^2}}{\frac{e^{\tau_0+\tau_1 x_i}}{1+e^{\tau_0+\tau_1 x_i}} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}\left(\frac{y_i-(\beta_0+\beta_1 x_i)}{\sigma_i}\right)^2} + \frac{1}{1+e^{\tau_0+\tau_1 x_i}} \mu_i e^{-\mu_i y_i} I_{(0,\infty)}(y_i)}$$

then $z_i = 1$ with probability h_i. In this way, the likelihood function for $\mathbf{z}$ is

$$L(\mathbf{z}/\mathbf{y}, \mathbf{x}, \mathbf{w}, \mathbf{\Theta}) = \prod_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{h_i^{z_i}(1-h_i)^{z_i}}$$
$$= \prod_{i=1}^{n} \frac{\left(\frac{e^{(\tau_0+\tau_1 x_i)}}{1+e^{(\tau_0+\tau_1 x_i)}} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}\left(\frac{y_i-(\beta_0+\beta_1 x_i)}{\sigma_i}\right)^2} \right)^{z_i} \left(\frac{1}{1+e^{\tau_0+\tau_1 x_i}} \mu_i e^{-\mu_i y_i} I_{(0,\infty)}(y_i) \right)^{1-z_i}}{\frac{e^{\tau_0+\tau_1 x_i}}{1+e^{\tau_0+\tau_1 x_i}} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}\left(\frac{y_i-(\beta_0+\beta_1 x_i)}{\sigma_i}\right)^2} + \frac{1}{1+e^{\tau_0+\tau_1 x_i}} \mu_i e^{-\mu_i y_i} I_{(0,\infty)}(y_i)} \tag{1.4}$$

and the likelihood function for $(\mathbf{y},\mathbf{z})$ is given by

$$L(\mathbf{y},\mathbf{z}/\mathbf{x},\mathbf{w},\mathbf{\Theta}) =$$

$$= \prod_{i=1}^{n} \frac{\left(\frac{e^{(\tau_0+\tau_1 x_i)}}{1+e^{(\tau_0+\tau_1 x_i)}}\frac{1}{\sqrt{2\pi\sigma_i}}e^{-\frac{1}{2}\left(\frac{y_i-(\beta_0+\beta_1 x_i)}{\sigma_i}\right)^2}\right)^{z_i}\left(\frac{1}{1+e^{\tau_0+\tau_1 x_i}}\mu_i e^{-\mu_i y_i} I_{(0,\infty)}(y_i)\right)^{1-z_i}}{\frac{e^{\tau_0+\tau_1 x_i}}{1+e^{\tau_0+\tau_1 x_i}}\frac{1}{\sqrt{2\pi\sigma_i}}e^{-\frac{1}{2}\left(\frac{y_i-(\beta_0+\beta_1 x_i)}{\sigma_i}\right)^2} + \frac{1}{1+e^{\tau_0+\tau_1 x_i}}\mu_i e^{-\mu_i y_i} I_{(0,\infty)}(y_i)}$$

$$\prod_{i=1}^{n}\left(\frac{e^{\tau_0+\tau_1 x_i}}{1+e^{\tau_0+\tau_1 x_i}}\frac{1}{\sqrt{2\pi\sigma_i}}e^{-\frac{1}{2}\left(\frac{y_i-(\beta_0+\beta_1 x_i)}{\sigma_i}\right)^2} + \frac{1}{1+e^{\tau_0+\tau_1 x_i}}\mu_i e^{-\mu_i y_i} I_{(0,\infty)}(y_i)\right)$$

$$= \prod_{i=1}^{n}\left(\frac{e^{(\tau_0+\tau_1 x_i)}}{1+e^{(\tau_0+\tau_1 x_i)}}\frac{1}{\sqrt{2\pi\sigma_i}}e^{-\frac{1}{2}\left(\frac{y_i-(\beta_0+\beta_1 x_i)}{\sigma_i}\right)^2}\right)^{z_i}\left(\frac{1}{1+e^{\tau_0+\tau_1 x_i}}\mu_i e^{-\mu_i y_i} I_{(0,\infty)}(y_i)\right)^{1-z_i}$$

For the prior distributions of the parameters we assume normal distributions with diagonal variance-covariance matrices and high variance values, so as to make the distributions approximately non informative. For simplicity, we suppose independence among the parameters a priori. These a priori distributions are given by,

$$\begin{aligned} \boldsymbol{\beta} &\sim N(\mathbf{b}, B^{-1}) \\ \boldsymbol{\gamma} &\sim N(\mathbf{g}, G^{-1}) \\ \boldsymbol{\lambda} &\sim N(\mathbf{l}, L^{-1}) \\ \boldsymbol{\tau} &\sim N(\mathbf{t}, T^{-1}) \end{aligned} \tag{1.5}$$

Where $B^{-1}, B^{-1}, G^{-1}, L^{-1}$ and T^{-1} have the general form $10^k I$ where I is the identity matrix whose dimension is that of the corrsponding vector of parameters. From (2.5) and (2.9) and using Bayes' theorem, the posterior distribution is given by

$$\pi(\Theta/\mathbf{x},\mathbf{w},\mathbf{y},\mathbf{z}) \propto p(\beta)\mathbf{p}(\gamma)\mathbf{p}(\lambda)\mathbf{p}(\tau)\mathbf{L}(\mathbf{y},\mathbf{z}/\mathbf{x},\mathbf{w},\mathbf{\Theta})$$

Thus, the full conditional posterior distributions for the parameters needed for the Gibbs sampling algorithm have the following forms:

$$\pi\left(\boldsymbol{\beta}\,|\mathbf{y},\mathbf{x},\mathbf{w},\boldsymbol{\lambda},\boldsymbol{\tau},\boldsymbol{\gamma}\right)\propto$$
$$\frac{|B|^{\frac{1}{2}}}{\left(\sqrt{2\pi}\right)^{2}}\exp\left(-\frac{1}{2}\left(\boldsymbol{\beta}-\mathbf{b}\right)'B\left(\boldsymbol{\beta}-\mathbf{b}\right)\right)\exp\left(-\frac{1}{2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)'\Sigma_z^{-1}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)\right) \tag{1.6}$$

where $\mathbf{X} = [1\ \mathbf{x}]$, $\Sigma_z^{-1}=$ diag $(z_1/\sigma_1^2,\ldots,z_n/\sigma_n^2)$, $\sigma_i^2=$ exp $(\gamma_0+\gamma_1 w_i)$, so $\beta\sim N(\mu,\Sigma)$ with $\Sigma^{-1}=B+\mathbf{X}'\Sigma_z^{-1}\mathbf{X}, \quad \Sigma^{-1}\mu = Bb+\mathbf{X}'\Sigma_z^{-1}\mathbf{y}$

$$\pi\left(\boldsymbol{\lambda}|\mathbf{y},\mathbf{x},\mathbf{w},\boldsymbol{\beta},\boldsymbol{\tau},\boldsymbol{\gamma}\right)\propto$$
$$\frac{|L|^{\frac{1}{2}}}{(\sqrt{2\pi})^{2}}\exp\left(-\frac{1}{2}(\boldsymbol{\lambda}-\mathbf{l})'L(\boldsymbol{\lambda}-\mathbf{l})\right)\frac{\exp\left(-\sum_{i=1}^{n}\frac{(1-z_i)y_i}{\exp((\lambda_0+\lambda_1 x_i))}\right)}{\prod_{i=1}^{n}\exp(\lambda_0+\lambda_1 x_i)(1-z_i)} \tag{1.7}$$

$$\pi\left(\boldsymbol{\tau}|\mathbf{y},\mathbf{x},\mathbf{w},\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\gamma}\right)\propto$$
$$\frac{|T|^{\frac{1}{2}}}{(\sqrt{2\pi})^{2}}\exp\left(-\frac{1}{2}\left(\boldsymbol{\tau}-\mathbf{t}\right)'T\left(\boldsymbol{\tau}-\mathbf{t}\right)\right)\frac{\exp(\sum_{i=1}^{n}z_i(\tau_0+\tau_1 x_i))}{\prod_{i=1}^{n}(1+\exp(\tau_0+\tau_1 x_i))} \tag{1.8}$$

$$\pi\left(\boldsymbol{\gamma}\,|\mathbf{y},\mathbf{x},\mathbf{w},\boldsymbol{\lambda},\boldsymbol{\tau},\boldsymbol{\beta}\right)\propto$$
$$\frac{|G|^{\frac{1}{2}}}{\left(\sqrt{2\pi}\right)^{2}}\exp\left(-\frac{1}{2}\left(\boldsymbol{\gamma}-\mathbf{g}\right)'G\left(\boldsymbol{\gamma}-\mathbf{g}\right)\right)\frac{\exp\left(-\frac{1}{2}\sum_{i=1}^{n}z_i\frac{(y_i-\beta_0-\beta_1 x_i)^2}{\sigma_i^2}\right)}{\left(\prod_{i=1}^{n}(\sigma_i^2)^{z_i}\right)^{\frac{1}{2}}} \tag{1.9}$$

Observe that $\pi(\boldsymbol{\beta}|\mathbf{y},\mathbf{x},\mathbf{w},\boldsymbol{\lambda},\boldsymbol{\tau},\boldsymbol{\gamma})$ is a density which can be sampled from, so the Gibbs sampler will be used for β. For the other parameters, the Metropolis Hastings algorithm will be applied.

Based on the ideas in Cepeda and Gamerman (2001), the transition kernel

q for $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$ is a normal density which is obtained approximating the distribution of a working observation to a normal, and combining it with the prior distribution of the parameter. To be precise, it follows the steps:

- For γ : The following analysis is based on the fact that the observations y_i involved in the conditional distribution for γ are normal. Let $t = (y-\beta_0-\beta_1 x)^2$, then $E(t) = \sigma^2$, $\mathrm{Var}(t) = 2\sigma^4$, where σ^2=exp$(\gamma_0+\gamma_1 w)$, since $\frac{(y-\beta_0-\beta_1 x)^2}{\sigma^2} \sim \chi^2(1)$. If $g(t)$ =ln(t), we use a first order Taylor approximation for g around σ^2=exp$(\gamma_0+\gamma_1 w)$ to get

$$\begin{aligned} g(t) \approx g(\sigma^2) + g'(\sigma^2)(t-\sigma^2) &= \gamma_0 + \gamma_1 w + \frac{1}{\sigma^2}(t-\sigma^2) \\ &= \gamma_0 + \gamma_1 w + \frac{(y-\beta_0-\beta_1 x)^2}{\sigma^2} - 1 \end{aligned} \tag{1.10}$$

  Defining the working observation ỹ as ỹ $= \gamma_0+\gamma_1 w+\frac{(y-\beta_0-\beta_1 x)^2}{\sigma^2}-1$, we have E(ỹ) $=\gamma_0+\gamma_1 w$, and Var(ỹ)=2. Assuming ỹ $\sim N(\gamma_0+\gamma_1 w, 2)$, and using the prior for γ, $N(\mathbf{g}, G^{-1})$, the posterior distribution $\tilde{\pi}(\boldsymbol{\gamma})$ is obtained:

$$\begin{aligned} &\tilde{\pi}(\boldsymbol{\gamma}) \propto \\ &\frac{|G|^{\frac{1}{2}}}{(\sqrt{2\pi})^2}\left(-\frac{1}{2}(\boldsymbol{\gamma}-\mathbf{g})')G(\boldsymbol{\gamma}-\mathbf{g})\right)\frac{|2I|^{-\frac{1}{2}}}{(\sqrt{2\pi})^n}\exp\left(-\frac{1}{2}(\tilde{y}-\mathbf{W}\boldsymbol{\gamma})'(2I)^{-1}(\tilde{y}-\mathbf{W}\boldsymbol{\gamma})\right) \end{aligned} \tag{1.11}$$

  where $\mathbf{W} = [1\ \mathbf{w}]$
  That is $\tilde{\pi}(\boldsymbol{\gamma}) = N(\mu, \Sigma)$ where $\Sigma^{-1} = G + \frac{1}{2}\mathbf{W}'\mathbf{W}$, and $\Sigma^{-1}\mu = Gg + \frac{1}{2}\mathbf{W}'\tilde{y}$.
  So, a proposed kernel for $\boldsymbol{\gamma}$ is $q(\boldsymbol{\gamma}) = \tilde{\pi}(\boldsymbol{\gamma})$

- For $\boldsymbol{\lambda}$: The observations y_i in the conditional posterior for $\boldsymbol{\lambda}$ are exponentially distributed so the following analysis applies. Let $t = y$ where $y \sim$ exponential (α) and α=exp$(\lambda_0 + \lambda_1 x)$. Then E$(y) = \alpha$, Var$(y) = \alpha^2$. As before, let $g(t)$ =ln(t). Expanding around $E(y) = \alpha$ we obtain

$$g(t) \approx g(\alpha) + g'(\alpha)(t - \alpha)$$

Defining the working observation

$$\tilde{y} = g(\alpha) + g'(\alpha)(t - \alpha) = \lambda_0 + \lambda_1 x + \frac{y - \exp(\lambda_0 + \lambda_1 x)}{\exp(\lambda_0 + \lambda_1 x)},$$

then $E(\tilde{y}) = g(\alpha) = \lambda_0 + \lambda_1 x$, Var $(\tilde{y}) = (g'(\alpha))^2 \text{Var}(t) = \frac{1}{\alpha^2}\alpha^2 = 1$. Assuming $\tilde{y} \sim N(\lambda_0 + \lambda_1 x, 1)$, and using the prior for λ, $N(1, L^{-1})$, we obtain the posterior distribution for $\boldsymbol{\lambda}$

$$\tilde{\pi}(\boldsymbol{\lambda}) \propto \frac{|L|^{\frac{1}{2}}}{(\sqrt{2\pi})^2}\exp\left(-\frac{1}{2}(\boldsymbol{\lambda} - \mathbf{1})')L(\boldsymbol{\lambda} - \mathbf{1})\right) \frac{1}{(\sqrt{2\pi})^n}\exp\left(-\frac{1}{2}(\tilde{y} - \mathbf{X}\boldsymbol{\lambda})'(\tilde{y} - \mathbf{X}\boldsymbol{\lambda})\right)$$

That is $\tilde{\pi}(\boldsymbol{\lambda}) = N(\mu, \Sigma)$ where $\Sigma^{-1} = L + \mathbf{X}'\mathbf{X}$, and $\Sigma^{-1}\mu = L1 + \mathbf{X}'\tilde{y}$. So, a proposed kernel for $\boldsymbol{\lambda}$ is $q(\boldsymbol{\lambda}) = \tilde{\pi}(\boldsymbol{\lambda})$.

There are no y observations involved in the posterior for τ. In the simulations and the applications, a random walk was tried for q.
When the model is homoscedastic a conjugate prior or a locally uniform prior can be used to obtain a Gamma posterior distribution (Achcar et al.,1999a,b). In this case, the Gibbs sampling algorithm can be used to simulate the parameters of the variance.

1.3 A Simulation Study

In this section we show the results of four simulations. The algorithms were programmed in Matlab. In subsection 1.3.1 a summarizing table and the resulting chains from a first simulation are given. In subsection 1.3.2 the results of three simulations are summarized in corresponding tables, being an objective the comparison of the estimations of the parameters of both distributions. The parameters of both distributions were the same in the three simulations, but the parameters of the incidence probabilities changed to generate respectively quite the same number of observations from the normal and the exponential distributions, more from the normal than from the exponential, and more from the exponential than from the normal. For all cases we assumed approximately non-informative normal prior distributions of the form $N(\mathbf{a}, 10^2 I)$ were $\mathbf{a} = [0, 0]'$. The data for this simulation were generated as follows. 100 values of the variables X and W were generated from a uniform distribution $U(0, 10)$. Then, for each of the (x_i, w_i) an observation u from the $U(0, 1)$ is generated; if $u < a_1$, an observation y_i from the normal distribution is generated, else an observation from the exponential distribution is generated according to the true parameter values. The chains have 6.000 observations, and every tenth observation is chosen.

1.3.1 First Simulation

The first part of the computer program generates the $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ sample according to the true parameter values and the values $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ and $\boldsymbol{w} = (w_1, w_2, \ldots, w_n)$ of the explanatory variables $\boldsymbol{X}$ and $\boldsymbol{W}$, respectively. The second part estimates the parameters, and the third generates plots of the chains.

A first simulation is described as follows: values for the explanatory variables $\boldsymbol{X}$ and $\boldsymbol{W}$, are drawn from a uniform distribution in the interval [0,10], where 100 observations were generated for each variable. Depending on the value of τ, which is the parameter of the incidence probability $P(\mathbf{x}, \tau)$, the values of $\mathbf{y}$ will be generated from a normal distribution or from an exponential distribution, with the corresponding parameters. It is expected that if $P(\mathbf{x}, \tau) > 1/2$, that is, when $\tau_0 + \tau_1 x_i > 0$, the probability of being normal is greater than that of being exponential.

In table 1.1 there is a summary of the simulation results, where t.v. means true value, and s.d denotes the standard deviation.. Figures 1.1, 1.2, 1.3 and 1.4 show the behavior of the chains of the estimated parameters. The estimations given in table 1.1 are based on the last 500 observations of the chains. 95% credible intervals contain the true parameter values except for τ_0 and τ_1.

	β_0	β_1	γ_0	γ_1	τ_0	τ_1	λ_0	λ_1
t.v	7	1.6	-2	0.15	-0.2	0.03	0.4	0.15
Mean	6.97	1.60	-1.50	0.09	-0.69	0.12	0.32	0.17
S.d	0.17	0.03	0.46	0.07	0.29	0.06	0.26	0.05

Table 1.1: First simulation

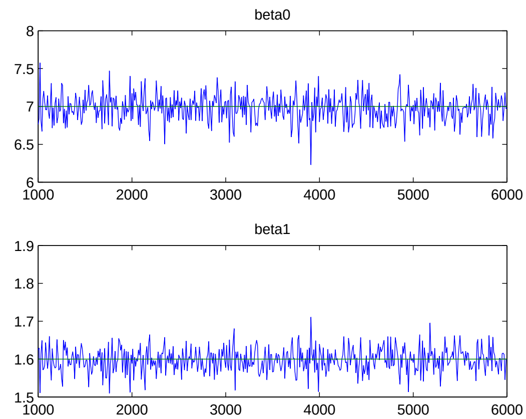

Figure 1.1: Chains for β, first simulation

1.3.2 Comparative simulations

Observe that there are three cases which make the incidence probabilities generate more, quite the same and less values, respectively from the normal than from the exponential distribution, that is, when $\tau_0 + \tau_1 x_i > (=, <) 0$. The values of x and w are drawn from a uniform distribution in the interval $[0, 10]$.

These runs are summarized in tables 1.2, 1.3, 1.4, where t.v means true value, b.e means bayesian estimate and se is the standard error. 95% credible intervals contain the true parameter values exept for τ_0 in the second run, where there are more obsevartions from the normal distribution.

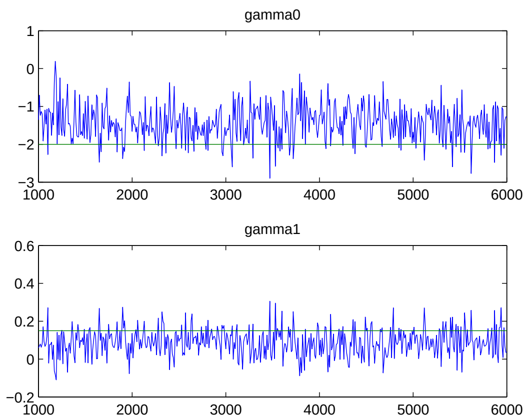

Figure 1.2: Chains for γ, first simulation

	β_0	β_1	γ_0	γ_1	τ_0	τ_1	λ_0	λ_1
t.v	1	0.5	-0.5	0.05	0	0	1	0.7
b.e	1.13	0.49	-0.22	0.05	-0.22	0.02	0.98	0.7
(se)	(0.20)	(0.034)	(0.3)	(0.056)	(0.29)	(0.04)	(0.18)	(0.03)

Table 1.2: Quite the same number of observations ($\tau_0 + \tau_1 x_i = 0$)

1.4 Applications

1.4.1 Insects

This example is based on a data set shown in Achcar et al (1999b). The variable y represents the number of insects dead after certain dose, x, of an insecticide is applied. In this example we take $w = x$, w being the covariate affecting the variance of the normal distribution. There are 317 observations in this data set. We assumed independent normal flat (large variance)

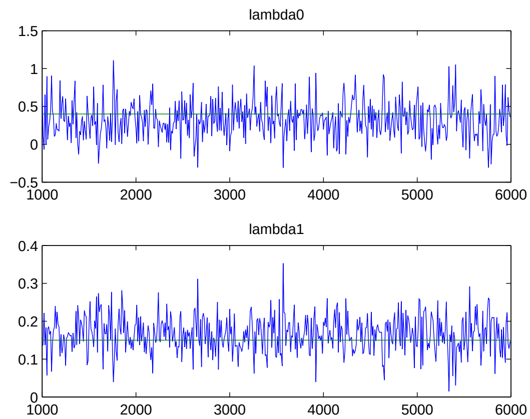

Figure 1.3: Chains for λ, first simulation

	β_0	β_1	γ_0	γ_1	τ_0	τ_1	λ_0	λ_1
t.v	1	0.5	-0.5	0.05	1	0.01	1	0.7
b.e	1.12	0.51	-0.6	0.076	0.31	0.15	0.8	0.73
(se)	(0.15)	(0.025)	(0.19)	(0.0352)	(0.36)	0.06	(0.23)	(0.04)

Table 1.3: More observations from the normal ($\tau_0 + \tau_1 x_i > 0$)

prior distributions for all parameters of the model, in order to express the absence of information about the parameters. The results are shown in table 1.5, with corresponding figures 1.5,1.6,1.7 and 1.8. In this example the value of $\tau_0 = 6.71$ shows that 99.88% of the observations correspond to the normal distribution and 0.12% correspond to the exponential , not enough ($0.9988 * 317 \approx 317$) to estimate λ_0 and λ_1, the parameters of the exponential distribution. A 95% credible interval for γ_1 is $[-2.0134, -1.4]$, showing that as x increases, the variance decreases, which agrees with the example, since it is expected that for large doses of the insecticide, the insects die homoge-

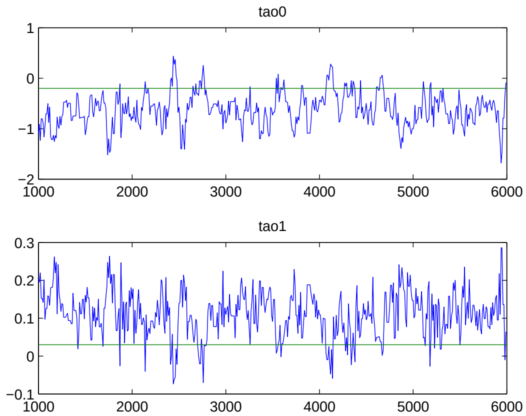

Figure 1.4: Chains for τ, first simulation

	β_0	β_1	γ_0	γ_1	τ_0	τ_1	λ_0	λ_1
t.v	1	0.5	-0.5	0.05	-1	-0.01	1	0.7
b.e	0.68	0.55	-0.72	0.04	-1.23	0.034	1.15	0.68
(se)	(0.25)	(0.04)	(0.45)	(0.08)	(0.37)	(0.05)	(0.15)	(0.02)

Table 1.4: More observations from the exponential ($\tau_0 + \tau_1 x_i < 0$)

nously fast. A 95% credible interval for β_1 also rejects the hypothesis $\beta_1 = 0$. The Jarque-Bera test for normality was applied to the distributions of the parameters, not rejecting normality, $\beta_1, \gamma_0, \gamma_1$.

	β_0	β_1	γ_0	γ_1	τ_0
	58.97	-140.95	8.59	-1.69	6.71
(se)	(12.32)	(16.16)	(0.20)	(0.18)	(1.37)

Table 1.5: Bayesian estimation of the parameters in the example of the insects

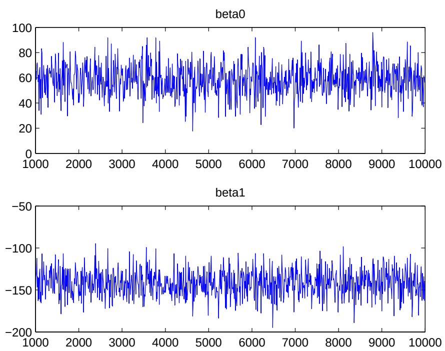

Figure 1.5: Chains for β, insects

1.4.2 DDT

This example is based on a data set consisting of 144 observations, shown in Mendenhall et al (1997). As in the former application, for the same reason, we also assume normal flat prior distributions for all the parameters in the model. In this example y represents the amount of DDT found in fishes at a distance $x = w$ from a contaminating plant beside a river. The results of the estimation are shown in table 1.6, and corresponding graphs are shown in 1.9, 1.10,1.11, 1.12, 1.13. The values of DDT (y) range from 0.11 to 1100, and the values of $x = w$ (position from the contaminating plant, being positive in the direction in which the river dies) range from -24 to 46. The large value of the mean of the exponential distribution, $\mu_e(x) = \exp(4.63 - 0.08x)$,

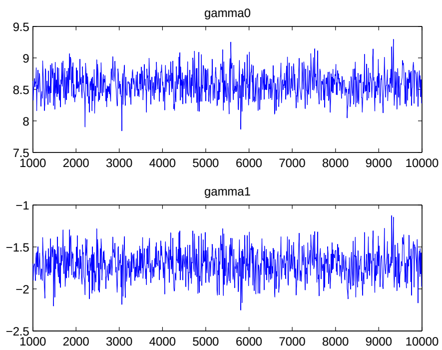

Figure 1.6: Chains for γ, insects

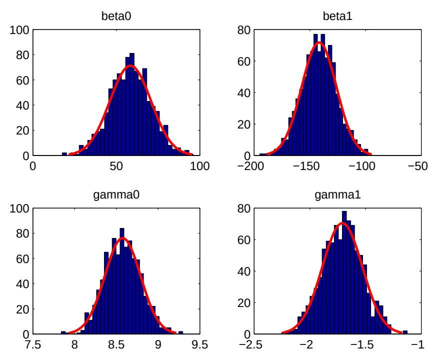

Figure 1.7: Histograms for the β and γ chains

suggests that the exponential distribution is modelling the large values of DDT, while the normal, with mean $\mu_n(x) = 5.84 + 0.13x$, seems to model smaller values of DDT. The value of $\tau_0 = 0.4603$, means that 61.31% of the observations are from the normal, and 38.69% are from the exponential distribution. All the parameters but τ_0 are statistically different from zero, according to 95% credible intervals. The Jarque-Bera normality test was applied to the chains of the estimated parameters, not rejecting normality

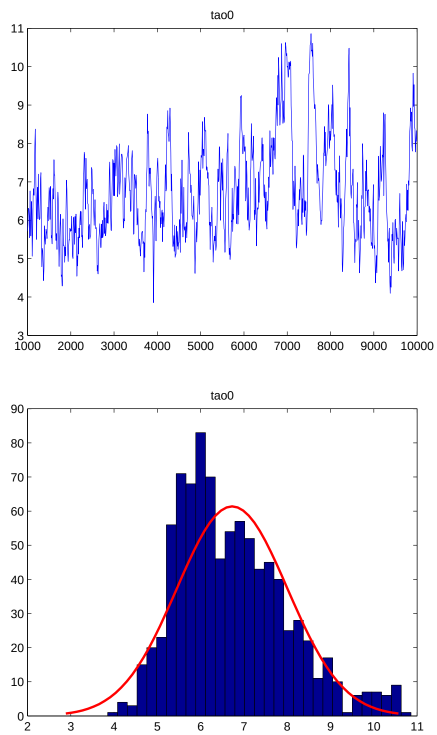

Figure 1.8: Chain and histogram for τ, insects

for $\beta_0, \beta_1, \gamma_1, \lambda_0, \lambda_1$. The Jarque-Bera Normality test is a form of Wald test where the null hypothesis is that the data are normal. The Jarque-Bera test statistic is defined in terms of sample estimates of the skewness and excess kurtosis based on a sample size n, and it is asymptotically chi-square with two degrees of freedom (Alexander, 2001)

1.5 Conclusions and Further Results

	β_0	β_1	γ_0	γ_1	τ_0	λ_0	λ_1
b.e	5.83	0.13	2.70	0.026	0.46	4.63	-0.08
(se)	(0.58)	(0.03)	(0.23)	(0.01)	(0.23)	(0.22)	0.01

Table 1.6: Bayesian estimation of the parameters in the example of DDT

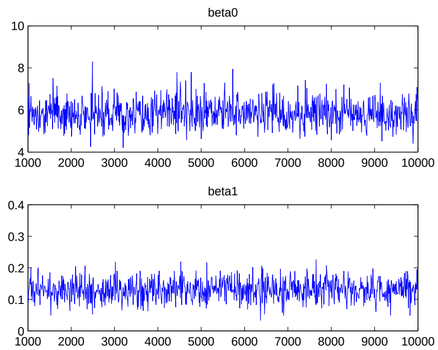

Figure 1.9: Chains for β, DDT

The results of the simulations showed good fits for the estimation of the parameters, in particular for the parameters of the variance of the normal distribution which were the innovation in the mixture model, allowing to explain heterogeneity, as was exemplified in the insects application. In this chapter we showed an example of a more general model in which there are k distributions in the mixture belonging to two different subfamilies of the biparametric exponential family, and the mean and variance of the distributions are modelled.

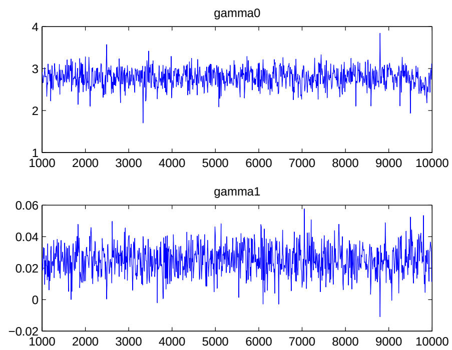

Figure 1.10: Chains for γ, DDT

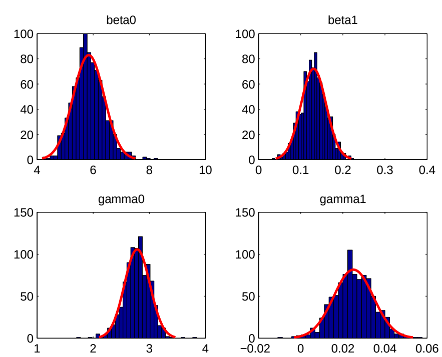

Figure 1.11: Histograms for β and γ, DDT

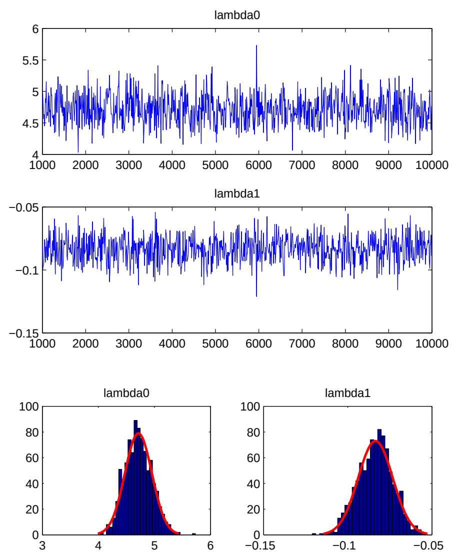

Figure 1.12: Chains and histograms for λ, DDT

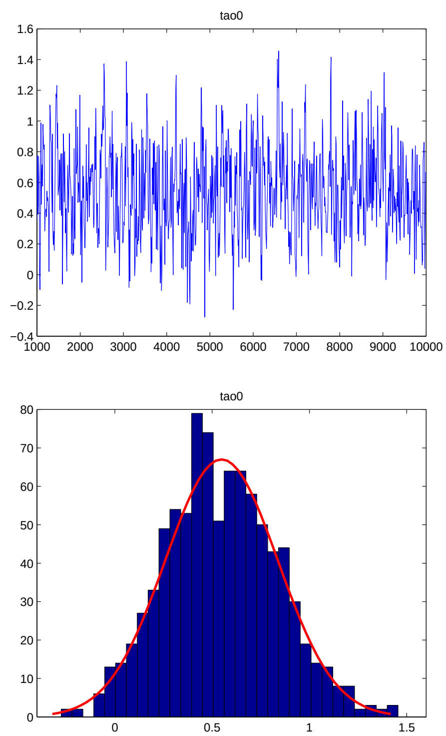

Figure 1.13: Chains and histograms for τ, DDT

Bibliography

[1] Achcar, J. A.; Pereira, G. A., (1999a). Mixture model for type II survival in the presence of covarites, Computacional Statistics, 14: 223-250

[2] Achcar, J. A.; Pereira, G. A., (1999b) Use of exponential power distribution for mixture models in the presence of covariates, Journal of Applied Statistics, Vol. 26, N°6, 669-679

[3] Alexander C.(2001) Market Models, a Guide to Financial Data Analysis, first edition, John Wiley & Sons.

[4] Carroll, J. R., Roeder K. and Wasserman L. (1996). Flexible parametric measurement error models. Technical Reports N 648, Department of Statistics, Carnegie Mellon University.

[5] Casella G. ; Mengersen K. L; Robert C. P. ; D. M. Titterington D. M.(2002) Perfect Sampler for Mixtures of distributions Journal of the Royal Statistics Society. Series B (Statistical Methodology), Vol. 64 N 4. pp 777-790

[6] Cepeda E.; Gamerman D (2001). Bayesian Modeling of Variance heterogeneity in Normal Regression Models. Brazilian Journal of Probability and Statistics, 14, pp. 207-221

[7] Cepeda E. and Gamerman D. (2005). Bayesian Methodology for modeling parameters in the two parameter exponential family. Estadstica, 57, 168 y 169, pp. 93 - 105.

[8] Chib, S. Greenberg, E (1995). the Metropolis-Hastings Algorithm. The American Statistician 49(4), 327-335.

[9] Dey, D. K., Kuo, L. and Sahu, S. K. (1995). A Bayesian predictive approach to determining the number of components in a mixture distribution. Statistics and Computing, 5, 297-305.

[10] Diebolt J.; Robert C. P. (1994)Estimation of Finite Mixture Distributions through Bayesian Sampling. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 56, N 2. pp. 363-375

[11] Gamerman D.; Lopes H.F. (2006) Markov Chain Monte Carlo, Taylor & Francis.

[12] Gelfand, A. E.; Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association, 85, 398-409.

[13] Mendenhall W.; Terry, S (1997) Probabilidad y Estadstica para Ingeniera y Ciencias, 4a edicin, Prentice Hall.

[14] Mengersen, K. and Robert, C. P. (1996). Testing for mixtures: A Bayesian entropic approach. In Bayesian Statistics 5 (eds. J. O. Berger, J. M. Bernardo, A. P. Dawid, D. V. Lindley A. F. M. Smith. Oxford University Press: London.

[15] Peng F.; Robert J.A. ; Tanner M (1996) A. Bayesian Inference in Mixtures-of-Experts and Hierarchical Mixture-of-Expert Model With an Application to Speech Recognition. Journal of the American Statistical Association, Vol. 91, N435. pp. 953-960.

[16] Polymenis, A. and Titterington, D. M. (1998). On the determination of the number of components in a mixture. Statistics and Probability Letter, 38, 295-298.

[17] Richarson S.; Green P. J.(1997) On Bayesian Analysis of Mixtures with an Unknown Number of Components. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 59, N 4. pp. 731-792.

[18] Robert, C. P. (1996). Mixtures of distributions: inference and estimation. In Practical Markov Chain Monte Carlo, (eds. W. R. Gilks, S. Richardson, D. J. Spiegelhalter). London: Chapman and Hall.

[19] Roeder K. and Wasserman L. (1997). Practical Bayesian density estimation using mixtures of normals. Journal of the American Statistical Association, 92, 894-902.

[20] Stephens, M. (1997). Bayesian methods for mixture of normal distributions. unpublished Ph.D. thesis, University of Oxford.

[21] Stephens, M. (1998). Bayesian analysis of mixtures with an unknown number of components - An alternative to reversible jump methods. 5

[22] Stephens, M. (2000). Dealing with label switching in mixture models. Journal of the Royal Statistical Society, B, 62, 795-809.

[23] Tanner M. A.; Wong W. H. (1987). The Calculation of Posterior Distributions by Data Augmentation. Journal of the American Statistical Association, Vol. 82, N398. pp. 528-540.

[24] Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). Statistical analysis of finite mixture distributions. Wiley: New York.

Chapter 2

MIXTURE OF DISTRIBUTIONS IN THE BIPARAMETRIC EXPONENTIAL FAMILY: A BAYESIAN APPROACH

Summary: In this chapter we propose mixture of distributions belonging to the biparametric exponential family, considering joint modelling of the mean and variance (or dispersion) parameters. As special cases we consider mixtures of normal and gamma distributions. A Bayesian methodology, using Markov Chain Monte Carlo (MCMC) methods, is proposed to obtain the posterior summaries of interest. We include simulations and real data examples to illustrate de performance of the proposal.

Key-words: mixture models, biparametric exponential family, Bayesian methods, MCMC simulation.

2.1 Introduction

Much has been written about mixtures of distributions, mainly distributions from the same subfamilies, modelling the means with covariates. Many ways to estimate the parameters have been used, among them maximum likelihood as in the EM algorithm, MCMC in Bayesian methodology using Gibbs samplers and Metropolis-Hastings algorithm algorithm and many more.

Peng et al. (1996) use mixture of normals in the presence of covariates. A general treatment for mixture models is given in Richardson and Green (1997) with an unknown number of distributions, all from the same subfamily. Achcar et al (1999a,b) mixes two types of distributions modelling the mean. Mazucheli and Achcar et al. (2001) in a model of survival time, model the error with a mixture of normal distributions, each having mean and variance constant. Wiper et al,(2001) work with a mixture of gamma distributions. Al-Saleh and Agarwal (2006) propose a finite mixture of gamma distributions but as a conjugate prior for the parameter of the Poisson distribution. F.Greselin and S. Ingrassia (2009) mix multivariate t distributions with constant means and variances, using a constrained EM algorithm. Yungtai Lo (2009) applies mixture of Normal distributions, modelling the mean with covariates and constant variance, in a study of Testosterone Deficiency in men.

Much has also been said about the number of components in the mixture.

Many have worked in the estimation of the number of components using different methods, Kass and Raftery (1995) work with Bayes factors, Mengersen and Robert (1996) use entropy distance, the known reversible jump MCMC of Richardson and Green (1997), Stephens (2000) works with birth and death processes. Dollena S. Hawkins et al (2001) determine the number of components in mixture of linear regression models using log likelihood measures. Nobile (2004) specifies the posterior distribution of the number of nonempty components under some constraints on the prior distribution.

Joint modelling of the mean and variance, not in mixtures, has already been proposed by Aitkin (1987), Cepeda and Gamerman (2001, 2004). In this chapter we study the mixture of k distributions belonging to two subfamilies of the biparametric exponential family, modelling both the mean and the variance.

If there is a one to one correspondence between the two usual parameters of the distribution of the family and the mean and the variance, any pair can be modelled to obtain the other pair, being more natural to model the mean and the variance.

For the biparametric exponential family, the density $f(x;\boldsymbol{\theta})$ will be expressed using the natural vector of parameters $\boldsymbol{\eta}$ as a function of the original parameters of the distribution $\boldsymbol{\theta}$; with this in mind, the density function has the form

$$f(x;\boldsymbol{\theta}) = \exp\left\{\boldsymbol{\eta}^t(\boldsymbol{\theta})T(x) - \xi(\boldsymbol{\theta})\right\} h(x),$$

where $\boldsymbol{\eta}^t(\boldsymbol{\theta}) = (\eta_1(\theta_1,\theta_2),\eta_2(\theta_1,\theta_2))$, $T^t(x) = (T_1(x),T_2(x))$. A system of two equations relating the mean and variance of $T(x)$ with the natural parameters of the family is given below, showing a general way to obtain these parameters from the mean and variance of $T(x)$, not necessarily being the easiest way to proceed in all the distributions of the family.

Differentiating with respect to θ_1 and θ_2 in the expression

$$\int_{\Re} \exp\left\{\boldsymbol{\eta}^t(\boldsymbol{\theta})T(x) - \xi(\boldsymbol{\theta})\right\} h(x)dx = 1$$

the moments of $T(x)$ can be obtained. In particular, the mean $E(T(x))$ and the variance $Var(T(x))$ can be solved from the set of two equations:

$$\begin{aligned} J'Var(T)J - J'E(T)D\xi - (D\xi)'E'(T)J + E(T_1)H_{\eta_1} + E(T_2)H_{\eta_2} \\ = H_\xi - (D\xi)'(D\xi) \qquad (2.1) \\ J'E(T) = (D\xi)' \end{aligned}$$

where $J = \left[\begin{array}{c} \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}} \end{array}\right]$, $D\xi = \left[\begin{array}{cc} \frac{\partial \xi}{\partial \theta_1} & \frac{\partial \xi}{\partial \theta_2} \end{array}\right]$, $H_{\eta_i}, i = 1, 2, H_\xi$ are the hessian marices of η_i and ξ with respect to $\boldsymbol{\theta}$. That is, if the mean and the variance of $T(X)$ are modelled, both parameters θ_1 and θ_2 are implicitely modelled.

Besides the introduction, the chapter has five more sections. In section 2 the formulation of the model is presented. In section 3 the proposed Bayesian methodology employed to fit the proposed model is described. In section 4 some results of simulated studies are presented. Section 5 includes results of a practical example. Finally, in section 6, some general conclusions are drawn.

2.2 The Model

Let Y be a random variable of interest and y_i, $i = 1, \ldots, n$, n independent realizations of Y with probability density function given by

$$f(y)=\sum_{j=1}^{k} a_j f_j(y;\boldsymbol{\theta}_j), \qquad (2.2)$$

where $f_j(y;\boldsymbol{\theta}_j)$, $j=1,...,k$, are distributions from the biparametric exponential family, and a_j are weights such that $0 \leq a_j \leq 1$, $a_1+\cdots+a_k=1$ and, in general, they are functions of the covariates. This model is called a "finite mixture distribution", the mixture weight a_j is the probability that the random variable y_i follows the distribution f_j, called a "mixture component." In this case, we assume that the mean and variance of each of the components of the mixture are given by regression models. Thus, particular cases could be the mixture of k normal distributions or k gamma distributions, where their means and variances are modelled as function of explanatory variables.

In what follows, the observations are $\mathbf{y}=(y_1,\ldots,y_n)'$ with covariate matrices $\mathbf{X}$ and $\mathbf{W}$, whose row vectors are $\mathbf{x}_i=(1,x_{i1},\ldots,x_{il})$, $\mathbf{w}_i=(1,w_{i1},\ldots,w_{ir})$, $i=1,\ldots,n$, respectively, and the general notation for the vectors of covariates is $\mathbf{x}:=(1,x_1,x_2,\ldots,x_l)$ and $\mathbf{w}:=(1,w_1,w_2,\ldots,w_r)$. $\boldsymbol{\theta}_j$ denotes the biparametric vector of parameters of the jth distribution in the mixture.

In order to obtain a simplification in the application of Bayesian methodology, latent Bernoulli variables, z_{ij}, $i=1,...,n$, $j=1,...,k$, with success probability given by (2.3), are introduced (see for example, Tanner and Wong, 1987; or Casela et al, 2002).

$$h_{ij}=\frac{a_j(\mathbf{x}_i)f_j(y_i;\boldsymbol{\theta}_j)}{\sum_{j=1}^{k}a_j(\mathbf{x}_i)f_j(y_i;\boldsymbol{\theta}_j)} \qquad (2.3)$$

Thus, taking into account these hidden variables and using $\{\mathbf{w}_i\}_{i=1}^{n}$ to denote the set of vectors $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$, the joint density function for $\mathbf{z} = (z_{11}, \ldots, z_{1k}, z_{21}, \ldots, z_{2k}, \ldots, z_{n1}, \ldots, z_{nk})$ conditional to $(\mathbf{y}, \{\mathbf{x}_i\}_{i=1}^n, \{\mathbf{w}_i\}_{i=1}^n, \{\boldsymbol{\theta}_j\}_{j=1}^k)$ is

$$
\begin{aligned}
f(\mathbf{z}|\mathbf{y}, \{\mathbf{x}_i\}_{i=1}^n, \{\mathbf{w}_i\}_{i=1}^n, \{\boldsymbol{\theta}_j\}_{j=1}^k) &= \prod_{i=1}^n \prod_{j=1}^k h_{ij}^{z_{ij}} (1-h_{ij})^{1-z_{ij}} \\
&= \prod_{i=1}^n \prod_{j=1}^k \frac{(a_j(\mathbf{x}_i) f_j(y_i; \boldsymbol{\theta}_j))^{z_{ij}} \left(\sum_{l=1, l\neq j}^k a_l(\mathbf{x}_i) f_l(y_i; \boldsymbol{\theta}_l) \right)^{1-z_{ij}}}{\sum_{l=1}^k a_l(\mathbf{x}_i) f_l(y_i; \boldsymbol{\theta}_l)} \qquad (2.4)
\end{aligned}
$$

and the joint density function for $(\mathbf{y}, \mathbf{z})$ conditional to $(\{\mathbf{x}_i\}_{i=1}^n, \{\mathbf{w}_i\}_{i=1}^n, \{\boldsymbol{\theta}_j\}_{j=1}^k)$ is given by,

$$
\begin{aligned}
f(\mathbf{y}, \mathbf{z}| \{\mathbf{x}_i\}_{i=1}^n, \{\mathbf{w}_i\}_{i=1}^n, \{\boldsymbol{\theta}_j\}_{j=1}^k) &= \\
&= \prod_{i=1}^n \prod_{j=1}^k \frac{(a_j(\mathbf{x}_i) f_j(y_i; \boldsymbol{\theta}_j))^{z_{ij}} \left(\sum_{l=1, l\neq j}^k a_l(\mathbf{x}_i) f_l(y_i; \boldsymbol{\theta}_l) \right)^{1-z_{ij}}}{\sum_{l=1}^k a_l(\mathbf{x}_i) f_l(y_i; \boldsymbol{\theta}_l)} \times \\
&\qquad \prod_{i=1}^n \sum_{j=1}^k a_j(\mathbf{x}_i) f_j(y_i; \boldsymbol{\theta}_j) \\
&= \prod_{i=1}^n \prod_{j=1}^k (a_j(\mathbf{x}_i) f_j(y_i; \boldsymbol{\theta}_j))^{z_{ij}} \left(\sum_{l=1, l\neq j}^k a_l(\mathbf{x}_i) f_l(y_i; \boldsymbol{\theta}_l) \right)^{1-z_{ij}} \qquad (2.5)
\end{aligned}
$$

To be clear, let us exemplify with the mixture of normal and gamma distributions, with weights from the logistic distribution. We consider the model where the mixture components and the weights are given by

$$
\begin{aligned}
f_j(y) &= \frac{1}{\sqrt{2\pi}\sigma_j} \exp\Big(-\frac{1}{2} \left(\frac{y - \mathbf{x}\boldsymbol{\beta}_j}{\sigma_j} \right)^2 \Big), \quad \sigma > 0 \qquad (2.6) \\
f_j(y) &= \frac{1}{\Gamma(\alpha_j)\delta_j^{\alpha_j}} y^{\alpha_j - 1} e^{-\frac{y}{\delta_j}} I_{(0,\infty)}(y), \quad \alpha > 0,\ \delta > 0 \qquad (2.7) \\
a_j &= \frac{e^{\mathbf{x}\boldsymbol{\tau}_j}}{1 + \sum_{l=1}^{l=k-1} e^{\mathbf{x}\boldsymbol{\tau}_l}} \qquad (2.8)
\end{aligned}
$$

where $j = 1, ..., k$ for (2.6) and (2.7); $j = 1, \ldots, k-1$ for (2.8), and $a_k = 1 - \sum_{j=1}^{k-1} a_j = \frac{1}{1+\sum_{l=1}^{l=k-1} e^{\mathbf{x}\boldsymbol{\tau}_l}}$. Related to these distributions, the mean and the variance in the normal distributions are modelled by $\mu^N(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$ and $\sigma^2(\mathbf{w}) = \exp(\mathbf{w}\boldsymbol{\gamma}^N)$, respectively. For the gamma distribution the variance will be modelled as the one for the normal distribution, that is, $\sigma^2(\mathbf{w}) = \exp(\mathbf{w}\boldsymbol{\gamma}^g)$, where the indexes $'g'$ and $'N'$ refer to the gamma and normal distributions. For the mean of the gamma distribution we consider two models, the first one given by $\mu(\mathbf{x}) = \mathbf{x}\boldsymbol{\lambda}$ and the second by $\mu(\mathbf{x}) = \exp(\mathbf{x}\boldsymbol{\lambda})$. To return to the usual parameters of the gamma distribution, the equations are $\delta(\mathbf{x}, \mathbf{w}) = \frac{\sigma^2(\mathbf{w})}{\mu(\mathbf{x})}$, $\alpha(\mathbf{x}, \mathbf{w}) = \frac{\mu^2(\mathbf{x})}{\sigma^2(\mathbf{w})}$ which are solved from the expressions for the mean and variance as functions of the original parameters of the distribution, $\mu = \alpha\delta$ and $\sigma^2 = \alpha\delta^2$.

2.3 Bayesian Methodology

To illustrate the Bayesian methodology proposed to fit the models of mixture distribution, in which mean and variance (or dispersion) parameters of each component are modelled as regression models, we assume the normal gamma mixture given by (2.6), (2.7) and (2.8).

2.3.1 Prior distributions

For the prior distributions of the parameters we assume normal distributions with diagonal variance-covariance matrices and high variance values, so as to introduce the least amount of possible information. For the sake of simplicity, we suppose a priori independence among the parameters. These a priori distributions are given by,

$$\begin{aligned}
\boldsymbol{\beta}_{i_N} &\sim N(\mathbf{b}_{i_N}, B_{i_N}^{-1}), \\
\boldsymbol{\gamma}_{i_d}^{d} &\sim N(\mathbf{g}_{i_d}^{d}, (G_{i_d}^{d})^{-1}), \quad d = N, g \\
\boldsymbol{\lambda}_{i_g} &\sim N(\mathbf{l}_{i_g}, L_{i_g}^{-1}) \\
\boldsymbol{\tau}_{j} &\sim N(\mathbf{t}_{j}, T_{j}^{-1}), j = 1, ..., k-1
\end{aligned} \tag{2.9}$$

where $i_d \in \{1, ..., m_d\}$, m_N is the number of normal distributions in the mixture, m_g is the number of gamma distributions in the mixture and $m_N + m_g = k$. The vectors of parameters for the mean and variance of the i_N-th normal distribution are $\boldsymbol{\beta}_{i_N}$ and $\boldsymbol{\gamma}_{i_N}^{N}$ respectively; the vectors of parameters for the mean and variance of the i_g-th Gamma distribution are $\boldsymbol{\lambda}_{i_g}$ and $\boldsymbol{\gamma}_{i_g}^{g}$, respectively.

From (2.5) and (2.9) and using Bayes' theorem, the posterior distribution is given by

$$\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}, \{\mathbf{x}_i\}_{i=1}^{n}, \{\mathbf{w}_i\}_{i=1}^{n}) \propto P(\boldsymbol{\theta}) L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}, \{\mathbf{x}_i\}_{i=1}^{n}, \{\mathbf{w}_i\}_{i=1}^{n}),$$

where $\boldsymbol{\theta}$ is the vector of all the parameters involved, that is,

$$\boldsymbol{\theta} = \left(\{\boldsymbol{\beta}'_{i_N}\}_{i_N=1}^{m_N}, \{\boldsymbol{\gamma}'_{i_N}\}_{i_N=1}^{m_N}, \left\{\boldsymbol{\lambda}'_{i_g}\right\}_{i_g=1}^{m_g}, \left\{\boldsymbol{\gamma}'_{i_g}\right\}_{i_g=1}^{m_g}, \{\boldsymbol{\tau}'_{j}\}_{j=1}^{k-1} \right)'$$

and $P(\boldsymbol{\theta})$ the joint prior density function. Thus, given that the posterior $\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}, \{\mathbf{x}_i\}_{i=1}^{n}, \{\mathbf{w}_i\}_{i=1}^{n})$ is analytically intractable and it is not easy to get samples from, we propose to get samples from the conditional posterior distributions using Gibbs samplers and the Metropolis Hastings algorithm.

2.3.2 Conditional Posterior Distributions

The conditional posterior distribution for the mean parameters $\boldsymbol{\beta}$ in any of the normal distributions in the mixture is given by

$$\pi\left(\boldsymbol{\beta}_{i_N}\left|\mathbf{y},\mathbf{z},\{\mathbf{x}_i\}_{i=1}^n,\{\mathbf{w}_i\}_{i=1}^n,\boldsymbol{\gamma}_{i_N}\right.\right)\propto\frac{|B_{i_N}|^{\frac{1}{2}}}{\left(\sqrt{2\pi}\right)^2}\exp\left(-\frac{1}{2}\left(\boldsymbol{\beta_{i_N}}-\mathbf{b}_{i_N}\right)'B_{i_N}\left(\boldsymbol{\beta_{i_N}}-\mathbf{b}_{i_N}\right)\right)$$
$$\times\exp\left(-\frac{1}{2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta_{i_N}}\right)'\Sigma_z^{-1}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta_{i_N}}\right)\right),\quad(2.10)$$

where $\mathbf{X}$ and $\mathbf{W}$ are the covariate matrices for the mean and the variance respectively, and $\Sigma_z^{-1}=$ diag $(z_{1i_N}/\sigma_1^2,\ldots,z_{ni_N}/\sigma_n^2)$, with $z_{ji_N}=1$ if y_j is an observation from the $i_N th$ normal distribution and $\sigma_j^2=\exp\left(\mathbf{W}\boldsymbol{\gamma}_{i_N}\right)_j$. Thus, these posterior conditional distributions are given by

$$\boldsymbol{\beta}_{i_N}\sim N(\boldsymbol{\mu},\Sigma)\qquad(2.11)$$

where $\Sigma^{-1}=B_{i_N}+\mathbf{X}'\Sigma_z^{-1}\mathbf{X}$ and $\Sigma^{-1}\boldsymbol{\mu}=B_{i_N}\mathbf{b}_{i_N}+\mathbf{X}'\Sigma_z^{-1}\mathbf{y}$.

The conditional posterior for the mean parameters $\boldsymbol{\lambda}$ in any of the gamma distributions in the mixture is given by

$$\pi\left(\boldsymbol{\lambda}_{i_g}\left|\mathbf{y},\mathbf{z},\{\mathbf{x}_i\}_{i=1}^n,\{\mathbf{w}_i\}_{i=1}^n,\boldsymbol{\gamma}_{i_g}\right.\right)\propto$$
$$\frac{|L|^{\frac{1}{2}}}{(\sqrt{2\pi})^2}\exp\left(-\frac{1}{2}(\boldsymbol{\lambda}_{i_g}-\mathbf{l})'L(\boldsymbol{\lambda}_{i_g}-\mathbf{l})\right)\prod_{j\in J}\frac{1}{\Gamma(\alpha_j)\delta_j^{\alpha_j}}y_j^{\alpha_j-1}e^{-\frac{y_j}{\delta_j}}I_{(0,\infty)}(y_j),$$
$$(2.12)$$

where $J=\left\{j\left|j=1,...,n,z_{ji_g}=1\right.\right\}$, $\alpha_j=\alpha_j(\mu_j,\sigma_j^2)$, $\delta_j=\delta_j(\mu_j,\sigma_j^2)$, $\mu_j=\exp(\mathbf{X}\boldsymbol{\lambda}_{i_g})_j$ and $\sigma_j^2=\exp(\mathbf{W}\boldsymbol{\gamma}_{i_g})_j$.

The conditional posterior for $\boldsymbol{\tau}_r$, the vector of parameters of the weight a_r, $r = 1, \ldots, k-1$, is given by

$$\pi\left(\boldsymbol{\tau}_r \,\middle|\, \mathbf{y}, \mathbf{z}, \{\mathbf{x}_i\}_{i=1}^{n}, \{\boldsymbol{\tau}_l\}_{l=1, l\neq r}^{k-1}\right) \propto \frac{|T_r|^{\frac{1}{2}}}{(\sqrt{2\pi})^2} \exp\left(-\frac{1}{2}\left(\boldsymbol{\tau}_r - \mathbf{t}_r\right)' T_r \left(\boldsymbol{\tau}_r - \mathbf{t}_r\right)\right) \frac{\prod_{j\in J} \exp(\mathbf{X}\boldsymbol{\tau}_r)_j}{\prod_{i=1}^{n}\left(1 + \sum_{l=1}^{k-1}\exp(\mathbf{X}\boldsymbol{\tau}_l)_i\right)}, \tag{2.13}$$

where $J = \{j\,|j = 1, ..., n, z_{jr} = 1\}$.

The conditional posterior for $\boldsymbol{\gamma}_{i_N}$, the vector of parameters of the variance of the i_Nth normal distribution in the mixture, is given by

$$\pi\left(\boldsymbol{\gamma}_{i_N} \,\middle|\mathbf{y}, \mathbf{z}, \{\mathbf{x}_i\}_{i=1}^{n}, \{\mathbf{w}_i\}_{i=1}^{n}, \boldsymbol{\beta}_{i_N}\right) \propto \frac{|G_{i_N}|^{\frac{1}{2}}}{\left(\sqrt{2\pi}\right)^2} \exp\left(-\frac{1}{2}\left(\boldsymbol{\gamma}_{i_N} - \mathbf{g}_{i_N}\right)' G_{i_N}\left(\boldsymbol{\gamma}_{i_N} - \mathbf{g}_{i_N}\right)\right) \prod_{j\in J} \frac{\exp\left(-\frac{1}{2}\frac{\left(y_j - \left(\mathbf{X}\boldsymbol{\beta}_{\boldsymbol{i_N}}\right)_j\right)^2}{\sigma_j^2}\right)}{\left(\sigma_j^2\right)^{\frac{1}{2}}}, \tag{2.14}$$

where $J = \{j\,|j = 1; \ldots, n, z_{ji_N} = 1\}$, $\sigma_j^2 = \exp\left(\mathbf{W}\boldsymbol{\gamma}_{i_N}\right)_j$.

The conditional posterior distribution for $\boldsymbol{\gamma}_{i_g}$, the vector of parameters of the variance of the i_gth gamma distribution in the mixture, is given by

$$\pi\left(\boldsymbol{\gamma}_{i_g}|\mathbf{y}, \mathbf{z}, \{\mathbf{x}_i\}_{i=1}^{n}, \{\mathbf{w}_i\}_{i=1}^{n}, \boldsymbol{\lambda}\right) \propto \frac{|G_{i_g}|^{\frac{1}{2}}}{(\sqrt{2\pi})^2} \exp\left(-\frac{1}{2}(\boldsymbol{\gamma}_{i_g} - \mathbf{l}_{i_g})' G_{i_g}(\boldsymbol{\gamma}_{i_g} - \mathbf{l}_{i_g})\right) \prod_{j\in J} \frac{1}{\Gamma(\alpha_j)\delta_j^{\alpha_j}} y_j^{\alpha_j - 1} e^{-\frac{y_j}{\delta_j}} I_{(0,\infty)}(y_j), \tag{2.15}$$

where $J = \left\{j\,\middle|j = 1, ..., n, z_{ji_g} = 1\right\}$, $\alpha_j = \alpha_j(\mu_j, \sigma_j^2)$, $\delta_j = \delta_j(\mu_j, \sigma_j^2)$, $\mu_j = \exp(\mathbf{X}\boldsymbol{\lambda}_{i_g})_j$ and $\sigma_j^2 = \exp(\mathbf{W}\boldsymbol{\gamma}_{i_g})_j$.

2.3.3 Proposed Bayesian methodology

From (2.11) we can see that samples of $\boldsymbol{\beta}$ can be drawn directly from all its conditional posterior distribution, applying the Gibbs Sampler algorithm (Geman and Geman, 1984). For the other parameters, the full conditional distribution is analytically intractable and it is not easy to generate samples from it. Thus, we propose bayesian methodology, applying the Metropolis Hastings algorithm, to obtain samples of the posterior distribution.

Based on the ideas in Cepeda and Gamerman (2001, 2005), a transition kernel q for $\boldsymbol{\lambda}$, $\boldsymbol{\gamma}^N$ and $\boldsymbol{\gamma}^g$ is proposed as the combination of the normal prior distribution of the parameter and the likelihood function resulting from the assumption that working observation variables, defined by first order Taylor approximation, have normal distributions. To be precise, if $\mu = g(\mathbf{x}\boldsymbol{\lambda})$ and $\sigma^2 = g(\mathbf{w}\boldsymbol{\gamma})$, the working observation variable will be defined as the linear approximation of $g^{-1}(t)$ around the mean, in the first case, and around the variance in the other, choosing a random variable t, for each case, in such a way that $E(t) = \mu$ in the first case, and $E(t) = \sigma^2$ in the other. In general, if δ is the parameter to be modelled, for example μ or σ^2, the working observation variables are defined as follows:

$$\tilde{y} := g^{-1}(\delta) + (g^{-1})'(\delta)\,(t-\delta)\,,$$

where t is a random variable such that $E(t) = \delta$. Thus $E(\tilde{y}) = g^{-1}(\delta)$ and $\mathrm{Var}(\tilde{y}) = \left((g^{-1})'\right)^2(\delta)\mathrm{Var}(t)$. With this in mind, we proceed to define the transition kernel q in the Metropolis-Hastings algorithm for $\boldsymbol{\lambda}$, $\boldsymbol{\gamma}^N$ and $\boldsymbol{\gamma}^g$.

1. Definition of the transition kernel for the vector of parameters, $\boldsymbol{\gamma}^N$, of

the variance of the normal distribution.

The following analysis is based on the fact that the random variables Y_i involved in the full conditional distribution for $\boldsymbol{\gamma}$ have normal distribution with mean $\mu = \mathbf{x}\boldsymbol{\beta}$ and variance $\sigma^2 = \exp(\mathbf{w}\boldsymbol{\gamma})$ where, for the sake of simplicity in the notation, $\mathbf{x}$ and $\mathbf{w}$ stand for the i-th rows in the matrices $\mathbf{X}$ and $\mathbf{W}$ corresponding to the i-th observation of Y.

To determine the corresponding working observation $\tilde{Y}_i$, that for the general case will be denoted $\tilde{Y}$, we define the random variable t as $t = (Y - \mathbf{x}\boldsymbol{\beta})^2$. Observe that $E(t) = \sigma^2$ and $\mathrm{Var}(t) = 2\sigma^4$, since $\frac{(Y-\mathbf{x}\boldsymbol{\beta})^2}{\sigma^2} \sim \chi^2(1)$. Thus

$$\begin{aligned} g^{-1}(t) &\approx g^{-1}(\sigma^2) + (g^{-1})'(\sigma^2)(t - \sigma^2) \\ &= \mathbf{w}\boldsymbol{\gamma} + \frac{1}{\sigma^2}(t - \sigma^2) \\ &= \mathbf{w}\boldsymbol{\gamma} + \frac{(Y - \mathbf{x}\boldsymbol{\beta})^2}{\sigma^2} - 1 \end{aligned} \tag{2.16}$$

Defining working observation $\tilde{\mathrm{Y}}$ as $\tilde{\mathrm{Y}} := \mathbf{w}\boldsymbol{\gamma} + \frac{(Y-\mathbf{x}\boldsymbol{\beta})^2}{\sigma^2} - 1$, we have E($\tilde{\mathrm{Y}}$) $=\mathbf{w}\boldsymbol{\gamma}$, and Var($\tilde{\mathrm{Y}}$)=2. Assuming that $\tilde{Y}$ has a normal distribution $N(\mathbf{w}\boldsymbol{\gamma}, 2)$, the joint distribution of $\tilde{\mathbf{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)'$ is a multivariate normal distribution $N(\mathbf{W}\boldsymbol{\gamma}, 2I)$ where I is the $n \times n$ identity matrix. Thus, a transition kernel $q = \tilde{\pi}(\boldsymbol{\gamma})$ is obtained by combining this multivariate normal distribution with the prior normal distribution of $\boldsymbol{\gamma}$, applying the Bayes theorem (see equation 2.17).

$$\tilde{\pi}(\boldsymbol{\gamma}) \propto$$
$$\frac{|G|^{\frac{1}{2}}}{(\sqrt{2\pi})^2}\left(-\frac{1}{2}(\boldsymbol{\gamma}-\mathbf{g})')G(\gamma-\mathbf{g})\right)\frac{|2I|^{-\frac{1}{2}}}{(\sqrt{2\pi})^n}\exp\left(-\frac{1}{2}(\tilde{\mathbf{y}}-\mathbf{W}\boldsymbol{\gamma})'(2I)^{-1}(\tilde{\mathbf{y}}-\mathbf{W}\boldsymbol{\gamma})\right) \tag{2.17}$$

So, the transition kernel for $\boldsymbol{\gamma}$ is $q = N(\boldsymbol{\mu}, \Sigma)$, where

$$\Sigma^{-1} = G + \frac{1}{2}\mathbf{W}'\mathbf{W}$$
$$\Sigma^{-1}\boldsymbol{\mu} = G\mathbf{g} + \frac{1}{2}\mathbf{W}'\tilde{\mathbf{y}}$$

2. Definition of the transition kernel for the vector of parameters, $\boldsymbol{\lambda}$, of the mean of the gamma distribution.

   The following analysis is based on the fact that the observations Y_i involved in the full conditional distribution for λ have gamma distribution $\Gamma(\alpha, \delta)$ with $\mu_g = \alpha\delta$ and $\sigma^2 = \alpha\delta^2$.

   To define the working observation $\hat{Y}_i$ that for the general case will be denoted $\hat{Y}$, we set the random variable t as $t = Y$. Then, $E(t) = \mu_g$ and $\mathrm{Var}(t) = \sigma^2$, where σ^2=exp($\mathbf{w}\boldsymbol{\gamma}$) and $\mu_g = \exp(\mathbf{x}\boldsymbol{\lambda})$. The working observation is

$$\tilde{Y} := \mathbf{x}\boldsymbol{\lambda} + \left(\frac{Y - \exp\left(\mathbf{x}\boldsymbol{\lambda}\right)}{\exp\left(\mathbf{x}\boldsymbol{\lambda}\right)}\right),$$

   with, $E(\tilde{Y}) = \mathbf{x}\boldsymbol{\lambda}$ and $\mathrm{Var}(\tilde{Y}) = \frac{\sigma^2}{\mu_g^2} = \frac{\exp(\mathbf{w}\boldsymbol{\gamma})}{\exp(2\mathbf{x}\boldsymbol{\lambda})}$. Assuming that the distribution of $\tilde{\mathbf{Y}} = \left(\tilde{Y}_1, \ldots, \tilde{Y}_n\right)$ is a multivariate normal distribution

$N(\mathbf{X}\boldsymbol{\lambda}, A)$, where $A = \text{diag}\left(\frac{\exp(\mathbf{w}\boldsymbol{\gamma})_i}{\exp(2(\mathbf{x}\boldsymbol{\lambda})_i)}\right)$, and combining it with the prior for $\boldsymbol{\lambda}$, a posterior distribution $q = \tilde{\pi}(\boldsymbol{\lambda})$ is given by

$$\tilde{\pi}(\boldsymbol{\lambda}) \propto$$
$$\frac{|L|^{\frac{1}{2}}}{(\sqrt{2\pi})^2}\left(-\frac{1}{2}(\boldsymbol{\lambda}-\mathbf{l})')L(\boldsymbol{\lambda}-\mathbf{l})\right)\frac{|A|^{-\frac{1}{2}}}{(\sqrt{2\pi})^n}\exp\left(-\frac{1}{2}(\tilde{\mathbf{y}}-\mathbf{X}\boldsymbol{\lambda})'(A)^{-1}(\tilde{\mathbf{y}}-\mathbf{X}\boldsymbol{\lambda})\right). \tag{2.18}$$

Thus, a proposed kernel for $\boldsymbol{\lambda}$, is a multivariate normal distribution $N(\boldsymbol{\mu}, \Sigma)$ where

$$\Sigma^{-1} = L + \mathbf{X}'A^{-1}\mathbf{X}$$
$$\Sigma^{-1}\boldsymbol{\mu} = L\mathbf{l} + \mathbf{X}'A^{-1}\tilde{\mathbf{y}}$$

3. Definition of the transition kernel for the vector of parameters, $\boldsymbol{\gamma}^g$, of the variance of the gamma distribution.

   To determine the corresponding working observation $\tilde{Y}_i$, that for the general case will be denoted $\tilde{Y}$, we define the random variable t as $t = (Y - \mathbf{x}\boldsymbol{\lambda})^2$. Observe that $E(t) = \sigma^2$ and $\text{Var}(t) = 3\sigma^4 + 6\frac{\sigma^6}{\mu_g^2} - \sigma^4$, where σ^2=$\exp(\mathbf{w}\boldsymbol{\gamma})$ and $\mu_g = \exp(\mathbf{x}\boldsymbol{\lambda})$. Thus, the working observation is

$$\tilde{Y} := \mathbf{w}\boldsymbol{\gamma} + \left(\frac{t}{\sigma^2} - 1\right),$$

   for which, $E(\tilde{Y}) = \mathbf{w}\boldsymbol{\gamma}$ and $\text{Var}(\tilde{Y}) = \frac{\text{E}\left((Y-\mu_g)^4\right)}{\sigma^4} - 1 = 2 + 6\frac{\sigma^2}{\mu_g^2}$. Assuming that the joint distribution of $\tilde{\mathbf{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)'$ is a multivariate normal distribution $N(\mathbf{W}\boldsymbol{\gamma}, A)$, where $A = \text{diag}\left(2 + 6\frac{\exp(\mathbf{w}\boldsymbol{\gamma})_i}{\exp(2(\mathbf{x}\boldsymbol{\lambda})_i)}\right)$, and combining it with the prior for $\boldsymbol{\gamma}$, a posterior distribution $q = \tilde{\pi}(\boldsymbol{\gamma})$ is obtained (see equation 2.19).

$$\tilde{\pi}(\boldsymbol{\gamma}) \propto$$
$$\frac{|G|^{\frac{1}{2}}}{(\sqrt{2\pi})^2}\left(-\frac{1}{2}(\boldsymbol{\gamma}-\mathbf{g})')G(\boldsymbol{\gamma}-\mathbf{g})\right)\frac{|A|^{-\frac{1}{2}}}{(\sqrt{2\pi})^n}\exp\left(-\frac{1}{2}(\tilde{\mathbf{y}}-\mathbf{W}\boldsymbol{\gamma})'(A)^{-1}(\tilde{\mathbf{y}}-\mathbf{W}\boldsymbol{\gamma})\right) \tag{2.19}$$

That is, a proposed kernel for $\boldsymbol{\gamma}$, is a multivariate normal distribution $N(\boldsymbol{\mu}, \Sigma)$ where

$$\begin{aligned} \Sigma^{-1} &= G + \mathbf{W}'A^{-1}\mathbf{W} \\ \Sigma^{-1}\boldsymbol{\mu} &= G\mathbf{g} + \mathbf{W}'A^{-1}\tilde{\mathbf{y}} \end{aligned} \tag{2.20}$$

4. Samples of $\boldsymbol{\tau}$ are taken from a random walk.

2.4 Simulations

This section shows the results of three simulations. It includes the results of the cases in which the proposed Bayesian methodology was applied to fit mixture distribution models when the components of the mixture are normal and gamma distributions, with joint modelling of the mean and variance in all cases. In all the simulations independent normal flat prior distributions were assigned for the mean and variance parameters.

2.4.1 First simulated study

In the first simulation, the variable Y is explained from the mixture model $f(y) = a_1 f_N(y; x, w, \boldsymbol{\beta}, \boldsymbol{\gamma}_N) + a_2 f_G(y; x, w, \boldsymbol{\beta}, \boldsymbol{\gamma}_G)$ where f_N is as in (2.6), f_G as in (2.7) and a_1 as in (2.8), but with no covariates explaining the weights, that is, $a_1 = \frac{e^\tau}{1+e^\tau}$. The means for the normal and gamma distributions are respectively modeled by $\mu_N(x) = \beta_0 + \beta_1 x$ and $\mu_G = \lambda_0 + \lambda_1 x$. The variances are modeled by $\sigma_d^2(w) = exp(\gamma_0^d + \gamma_1^d w)$, where $d = N$ for the variance of the normal distribution, or $d = G$ for the variance of the gamma distribution. The data for this simulation were generated as follows. 300 values of the variables X and W were generated from a uniform distribution $U(0, 10)$. Then, for each of the (x_i, w_i) an observation u from the $U(0, 1)$ is generated; if $u < a_1$, an observation y_i from the normal distribution is generated, else an observation from the gamma distribution is generated according to the true parameter values.

To apply the Bayesian methodology, normal prior distributions of the form $N(\mathbf{a}, 10^2 I)$ were assigned to the parameters, where I stands for the identity 2×2 matrix, $\mathbf{a} = [1, 1]'$ for λ^G and $\mathbf{a} = [0, 0]'$ for the other parameters. True values and estimates with corresponding standard errors are given in table (2.1), and graphs and histograms of the chains are shown in figures (2.1) to (2.5).

The chains have 10.000 observations, and for the sake of independence, every tenth observation is chosen. Many chains were simulated starting from different initial values, eventually providing a rough indication of stationarity. The horizontal line in the graphs of the chains represent the true parameter values. The chains in the graphs are shown to begin in the observation

	β_0	β_1	γ_0^N	γ_1^N	τ_0	λ_0	λ_1	γ_0^g	γ_1^g
t.v	7	1.6	-0.7	0.15	0	2	0.15	1.5	-0.45
b.e	7.05	1.57	-0.92	0.18	-0.012	2.00	0.14	1.56	-0.47
(se)	(0.15)	(0.03)	(0.23)	(0.04)	(0.12)	(0.08)	(0.01)	(0.22)	(0.04)

Table 2.1: First simulation results, μ_G linear

1000, so, as can be seen, they show a quick convergence. The histograms of the chains do not indicate a great departure from normality, which was ratified using the Jarque-Bera normality test, not rejecting normality for almost all the parameters involved. As can be seen from the table, the bayesian estimations are very close to the true parameter values. Individual 95% credible intervals contain the true parameter values. The value given to the parameter of the incidence probability, $\tau_0 = 0$ makes $a_1 = 1/2$ meaning that half of the observations should go to the normal distribution. The graph of the chain of the probability of being gamma, figure (2.3), which mainly counts the percentage of observations going to the gamma distribution for each observation of the chain, corroborates the result.

2.4.2 Second simulated study

In this simulation the variables X, W and the general model for Y are as in the first simulation, with the same functional form for the mean of the Normal distribution and the variances of both distributions, but now the mean of the Gamma distribution is non linear and has the form $\mu_G(x) = \exp(\lambda_0 + \lambda_1 x)$. True values and estimates with corresponding standard errors are given in table (2.2), and graphs and histograms of the chains are shown in figures (2.6)

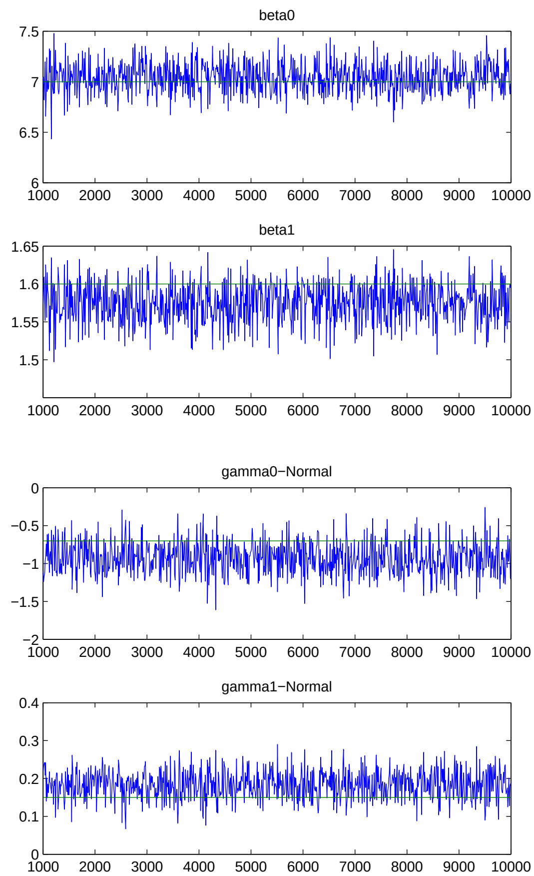

Figure 2.1: Chains for β and γ of the normal distribution, first simulation

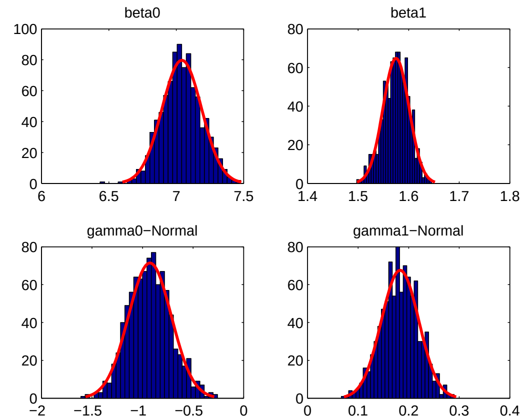

Figure 2.2: Histograms for β and γ of the normal distribution, first simulation

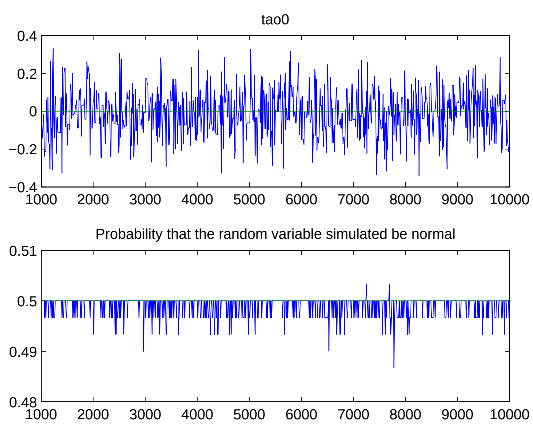

Figure 2.3: Chains for τ and probability of being Gamma,first simulation

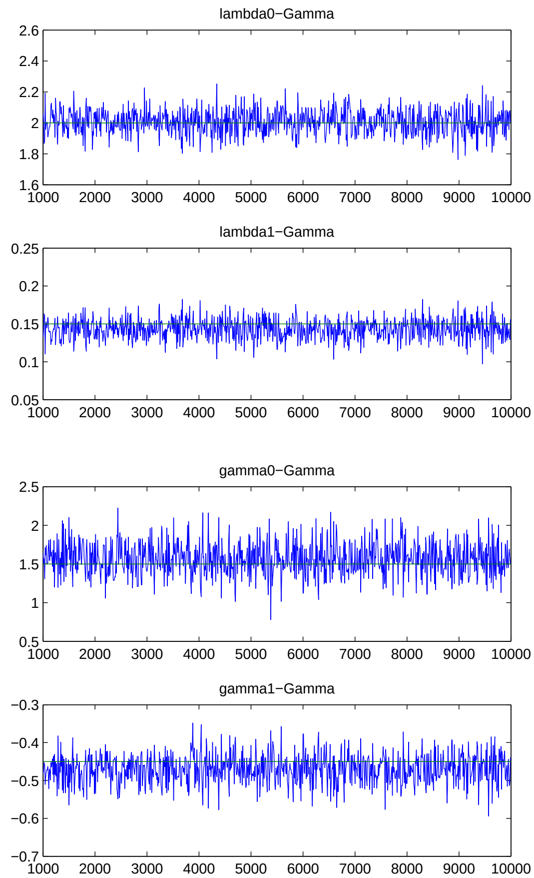

Figure 2.4: Chains for λ and γ from the gamma distribution, first simulation

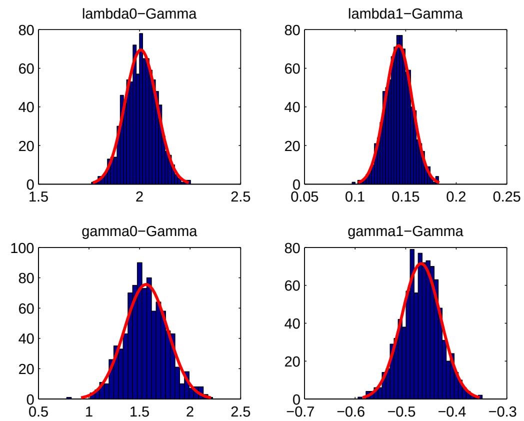

Figure 2.5: Histograms for the parameters of the Gamma distribution, first simulation

to (2.9). The chains have 10.000 observations, and every tenth observation is chosen. To apply Bayesian methodology, normal prior distributions of the form $N(\mathbf{a}, 10^2 I)$ were assigned to the parameters, where I stands for the identity 2×2 matrix, and $\mathbf{a} = [0, 0]'$ for all the parameters. As can be seen from the table, the bayesian estimations are very close to the true parameter values. Individual 95% credible intervals contain the true parameter values. A quick convergence can be seen in the graphs of the chains, and slight departures from normality can be seen in the graphs of the histograms, fact that was ratified for almost all the chain parameters, applying the Jarque-Bera normality test.

2.4.3 Third simulated study

The program used in the simulations runs for a general mixture model as described in section (2). To show this generality, in this simulation there is

	β_0	β_1	γ_0^N	γ_1^N	τ_0	λ_0	λ_1	γ_0^g	γ_1^g
t.v	-1.2	-0.08	0.2	0.05	0	1	0.05	0.16	0.0225
b.e	-1.003	-0.08	0.1	0.04	-0.07	0.99	0.05	0.12	0.06
(se)	(0.21)	(0.04)	(0.24)	(0.05)	(0.11)	(0.06)	(0.01)	(0.26)	(0.04)

Table 2.2: Second simulation results, μ_G non linear

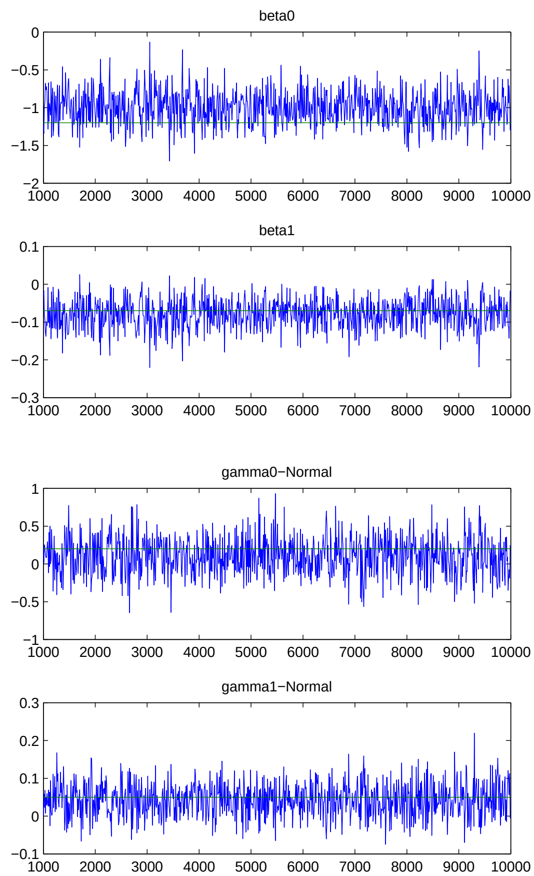

Figure 2.6: Chains for β and γ from the normal distribution, second simulation

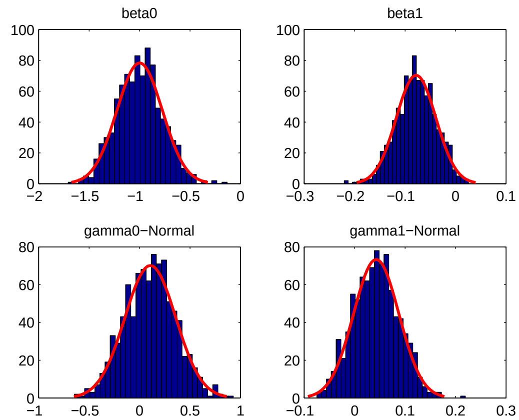

Figure 2.7: Histograms for β and γ, second simulation

a mixture of normal-gamma-normal distributions, with the mean of all the distributions explained by two covariates drawn from a uniform distribution $U(0, 10)$, the variance of the first Normal distribution constant, and the variances of the Gamma and the second Normal distribution explained by one covariate, drawn from a uniform distribution $U(0, 10)$. The models for the means and variances of the distributions are as in the first simulation, that is, all the means are linear, and the variances are exponential. The weights are $a_1 = \frac{\exp(\tau^{N_1})}{1+\exp(\tau^{N_1})+\exp(\tau^G)}$, $a_2 = \frac{\exp(\tau^G)}{1+\exp(\tau^{N_1})+\exp(\tau^G)}$, $a_3 = \frac{1}{1+\exp(\tau^{N_1})+\exp(\tau^G)}$. The notation for the parameters are as in the simulations above, in particular $\boldsymbol{\beta}_i^{N_j}$, $i = 0, 1, 2$ are the parameters of the mean of the j-th normal distribution, etc. To generate the observations from the mixture model a uniform$(0, 1)$ variable u is generated, if $u \leq a1$, an observation from the first normal distribution is generated, if $a1 < u \leq a1 + a2$ an observation from the gamma distribution is generated and if $a1 + a2 < u \leq 1$ an observation from the second normal distribution is generated. The results are shown in

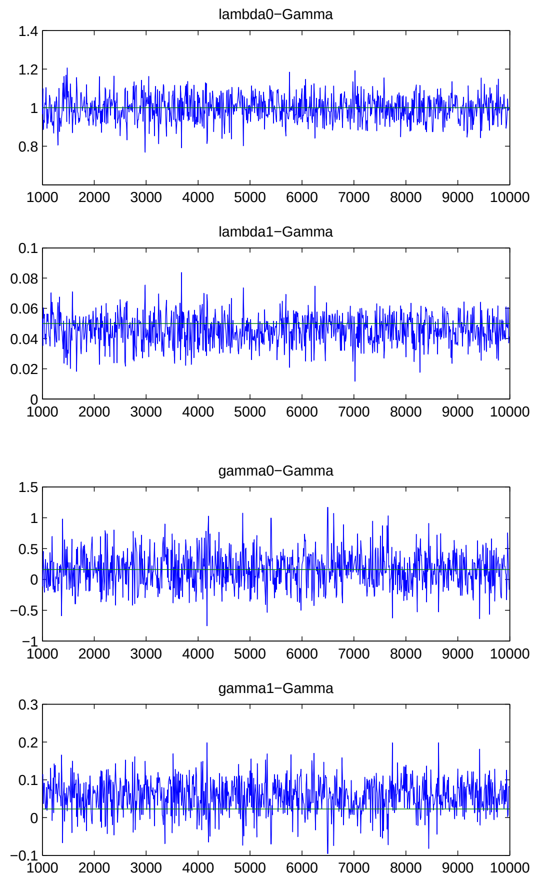

Figure 2.8: Chains for λ and γ from the gamma distribution, second simulation

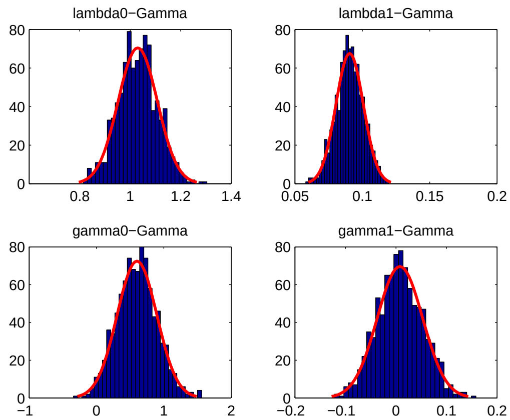

Figure 2.9: Histograms for λ and γ, μ_G non linear

table (2.3), and chains and histograms from figures (2.10) to (2.15) in the appendix. The true values for the parameters of the weights were $\tau^{N_1} = 0$, $\tau^G = 0$ that is, the probability of being from any of the distributions involved in the mixture is the same , 1/3. The first part of the algorithm generated 500 Y observations from the true model with weights 0.3200, 0.3200, 0.3600 respectively. The Bayes estimations and corresponding standard errors were $\hat{\tau}^{N_1} = 0$, (0), $\hat{\tau}^G = -0.06$, (0.096). 95% credible intervals for all the parameters contain the true parameter values. The Jarque-Bera normality test was applied, rejecting normality τ^{N_1}, τ^G, $\gamma_0^{N_2}$ and λ_0^G.

	$\beta_0^{N_1}$	$\beta_1^{N_1}$	$\beta_2^{N_1}$	$\gamma_0^{N_1}$	
t.v.	-0.2	-0.3	0.5	0.2	
b.e.	-0.21	-0.298	0.48	0.14	
(s.e.)	(0.24)	(0.03)	(0.03)	(0.11)	
	λ_0^G	λ_1^G	λ_2^G	γ_0^G	γ_1^G
t.v.	6.5	0.3	0.25	0.2	-0.1
b.e.	6.63	0.3	0.21	0.25	-0.62
(s.e.)	(0.2)	(0.03)	(0.03)	(0.24)	(0.48)
	$\beta_0^{N_2}$	$\beta_1^{N_2}$	$\beta_2^{N_2}$	$\gamma_0^{N_2}$	$\gamma_1^{N_2}$
t.v.	11.5	1.5	-0.3	-3	0.15
b.e.	11.54	1.5	-0.3	-3.13	0.27
(s.e.)	(0.04)	(0.006)	(0.006)	(0.2)	(0.34)

Table 2.3: Third simulation results

2.5 Applications

2.5.1 Home valuations

In this example the data base is from a country in Central America with 30.000 registers corresponding to households, each register having many variables, among them home valuations representing Y, household expenditure representing $x = w$, and a variable representing the socioeconomic level, ranging from one, the wealthiest, to 6, the poorest. 300 observations were taken randomly from this data base, just from socioeconomic levels one and six, in the same proportion as in the original data base. The model estimated is a mixture of a normal and a gamma distribution, with the mean of the gamma distribution being non linear. The results of the estimation are shown in table (2.4), and graphs from the chains and histograms are in figures 2.16

to 2.20 in the appendix, the BIC value is BIC=3.22. In this example, the incidence probability for the Gamma distribution is $e^{\tau}/(1+e^{\tau})$, with an estimated tau, $\tau = 2.68$, which means that there are more observations for the Gamma distribution than for the Normal distribution. The values for the parameters of the means, $\mu_N = 0.25 + 1.02x$, $\mu_G = e^{-3.86+8.49x}$, suggest that lower valuations are modelled by the Gamma distribution, which coincides with the values in the sample, since there are 212 observations corresponding to the poorest socioeconomic level, and 88 corresponding to the wealthiest. 95% credible intervals for λ_1 and γ_1^G imply that home valuations and its variability for the poorest, depend positively on consumption, while for the wealthiest, 95% confidence intervals for β_1 and γ_1^N contain the zero value, not rejecting the hypothesis that for this socioeconomic class, home valuation is not tied to consumption.

	β_0	β_1	γ_0^N	γ_1^N	τ_0	λ_0	λ_1	γ_0^w	γ_1^w
b.e	0.25	1.02	-3.48	-1.99	2.68	-3.86	8.49	-8.38	18.89
(se)	(0.069)	(0.69)	(0.68)	(6.24)	(0.35)	(0.084)	(0.64)	(0.23)	(1.62)

Table 2.4: Home valuations

2.6 Conclusions

In this chapter the mean and the variance of the component disributions in a mixture model were jointly modelled with covariates, particularly applied to the mixture of normal and gamma distributions. The results of the simulations showed good behavior in the convergence of the chains of the parameters, and the method was succesfully applied to a practical example.

In the introduction, equations (2.2) show the relation between the natural parameters of the biparametric exponential family, and the mean and the variance of $T(x)$. In the section of the Bayesian methodology, a general method to define the kernel distribution of the parameters of the mean and the variance in the Metropolis-Hastings algorithm is explained, so, in this way, the method can be generalized to the distributions of the biparametric exponential family.

The method can be easily implemented to any other biparametric family of distributions, for example, to the biparametric extreme value distributions.

Bibliography

[1] Achcar, J. A.; Pereira, G. A., (1999a). Mixture model for type II survival in the presence of covarites, Computational Statistics, 14: 223-250

[2] Achcar, J. A.; Pereira, G. A., (1999b) Use of exponential power distribution for mixture models in the presence of covariates, Journal of Applied Statistics, Vol. 26, N°6, 669-679.

[3] Aitkin M. (1987). Modeling variance heterogeneity in Normal Regression using Glim. Applied Statistics, 36: 332-339.

[4] Alexander C.(2001) Market Models, a Guide to Financial Data Analysis, first edition, John Wiley & Sons.

[5] Al-Saleh J. A.; Agarwal S. K. (2006). Finite mixture of gamma distributions: A conjugate prior. Computational Statistics & Data Analysis 51 (2007) 4369-4378.

[6] Carroll, J. R., Roeder K. and Wasserman L. (1996). Flexible parametric measurement error models. Technical Reports N 648, Department of Statistics, Carnegie Mellon University.

[7] Casella G. ; Mengersen K. L; Robert C. P. ; D. M. Titterington D. M.(2002) Perfect Sampler for Mixtures of distributions. Journal of the Royal Statistics Society. Series B (Statistical Methodology), Vol. 64 N 4. pp 777-790

[8] Cepeda E.; Gamerman D (2001). Bayesian Modeling of Variance heterogeneity in Normal Regression Models. Brazilian Journal of Probability and Statistics, 14, pp. 207-221

[9] Cepeda E. and Gamerman D. (2005). Bayesian Methodology for modeling parameters in the two parameter exponential family. Estadstica, 57, 168 y 169, pp. 93 - 105.

[10] Chib, S. Greenberg, E (1995). the Metropolis-Hastings Algorithm. The American Statistician 49(4), 327-335.

[11] Dey, D. K., Kuo, L. and Sahu, S. K. (1995). A Bayesian predictive approach to determining the number of components in a mixture distribution. Statistics and Computing, 5, 297-305.

[12] Diebolt J.; Robert C. P. (1994)Estimation of Finite Mixture Distributions through Bayesian Sampling. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 56, N 2. pp. 363-375

[13] Gamerman D.; Lopes H.F. (2006) Markov Chain Monte Carlo, Taylor & Francis.

[14] Gelfand, A. E.; Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association, 85, 398-409.

[15] Greselin F., Ingrassia S. (2009). Constrained monotone EM algorithms for mixtures of multivariate t distributions. Stat Comput (2010) 20: 9-22.

[16] Hawkins D. S., Allen D. M., Stomberg A. J., (2001). Determining the number of components in mixtures of linear models. Computational Statistics & Data Analysis 38 (2001) 15-48.

[17] Kass, R. E., Raftery, A.E., 1995. Bayes factors. Journal of American Statistical Association. 90 (430), 773-795.

[18] Mazucheli J., Achcar J. A., Kass Robert, Carta R. (2001). Use of mixture of distributions for the error in regression models for lifetime data. Brazilian Journal of Probability and Statistics (2001), 15, pp. 155-168.

[19] Mendenhall W.; Terry, S (1997) Probabilidad y Estadstica para Ingeniera y Ciencias, 4a edicin, Prentice Hall.

[20] Mengersen, K. and Robert, C. P. (1996). Testing for mixtures: A Bayesian entropic approach. In Bayesian Statistics 5 (eds. J. O. Berger, J. M. Bernardo, A. P. Dawid, D. V. Lindley A. F. M. Smith. Oxford University Press: London.

[21] Nobile A. (2004). On the posterior distribution of the number of components in a finite mixture. The Annals of Statistics (2004), Vol. 32, No. 5, 2044-2073.

[22] Peng F.; Robert J.A. ; Tanner M (1996) A. Bayesian Inference in Mixtures-of-Experts and Hierarchical Mixture-of-Expert Model With an Application to Speech Recognition. Journal of the American Statistical Association, Vol. 91, N435. pp. 953-960.

[23] Polymenis, A. and Titterington, D. M. (1998). On the determination of the number of components in a mixture. Statistics and Probability Letter, 38, 295-298.

[24] Richarson S.; Green P. J.(1997) On Bayesian Analysis of Mixtures with an Unknown Number of Components. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 59, N 4. pp. 731-792.

[25] Robert, C. P. (1996). Mixtures of distributions: inference and estimation. In Practical Markov Chain Monte Carlo, (eds. W. R. Gilks, S. Richardson, D. J. Spiegelhalter). London: Chapman and Hall.

[26] Roeder K. and Wasserman L. (1997). Practical Bayesian density estimation using mixtures of normals. Journal of the American Statistical Association, 92, 894-902.

[27] Stephens, M. (1997). Bayesian methods for mixture of normal distributions. unpublished Ph.D. thesis, University of Oxford.

[28] Stephens, M. (1998). Bayesian analysis of mixtures with an unknown number of components - An alternative to reversible jump methods. 5

[29] Stephens, M. (2000). Dealing with label switching in mixture models. Journal of the Royal Statistical Society, B, 62, 795-809.

[30] Tanner M. A.; Wong W. H. (1987). The Calculation of Posterior Distributions by Data Augmentation. Journal of the American Statistical Association, Vol. 82, N398. pp. 528-540.

[31] Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). Statistical analysis of finite mixture distributions. Wiley: New York.

[32] Wiper M., Rios Insua D., Ruggeri F. (2001). Mixtures of Gamma Distributions With Applications. Journal of computational and Graphical Statistics, Volume 10, Number 3, Pages 440-454.

[33] Yungtai Lo (2009). Estimating Age-specific Prevalence of Testosterone Deficiency in Men Using Normal Mixture Models. Journal of Data Science 7, 203-217.

2.7 Appendix: Graphs and histograms of the parameters in the third simulation and the home valuations example*

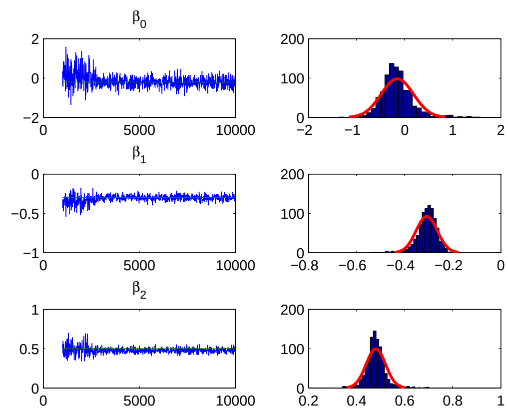

Figure 2.10: Chains and histograms for $\boldsymbol{\beta}^{\boldsymbol{N_1}}$, third simulation

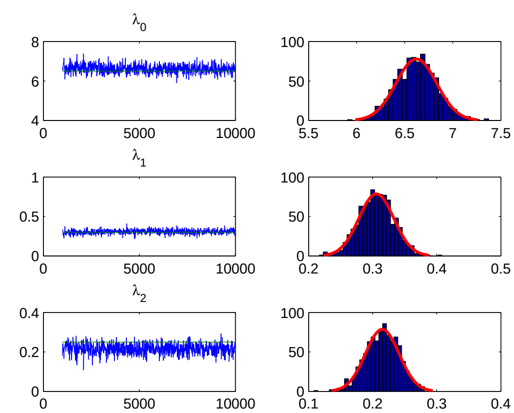

Figure 2.11: Chains and histograms for $\boldsymbol{\lambda}^{\boldsymbol{G}}$, third simulation

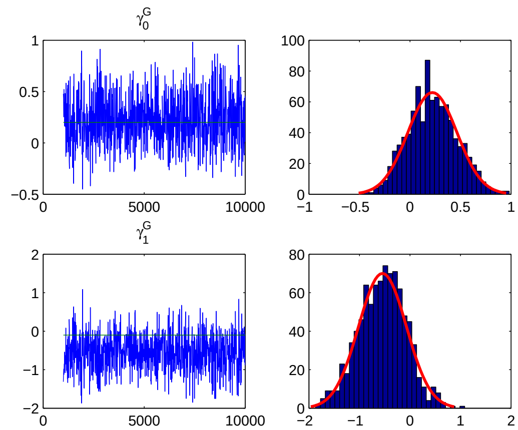

Figure 2.12: Chains and histograms for $\boldsymbol{\gamma}^{\boldsymbol{G}}$, third simulation

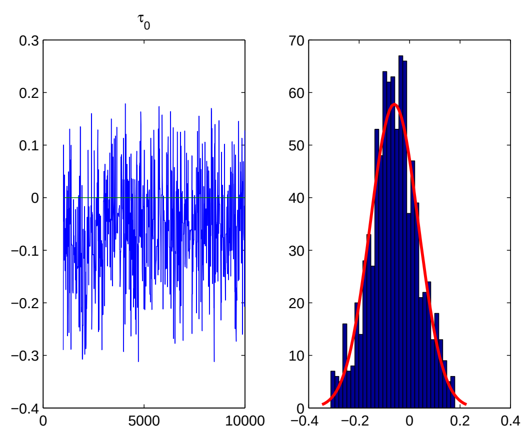

Figure 2.13: Chain and histogram for $\boldsymbol{\tau}^{\boldsymbol{G}}$, third simulation

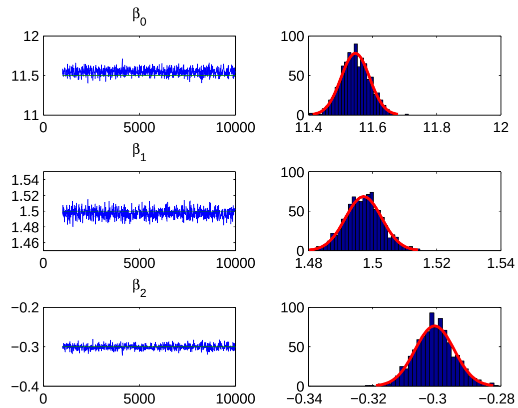

Figure 2.14: Chains and histograms for $\boldsymbol{\beta}^{\boldsymbol{N_2}}$, third simulation

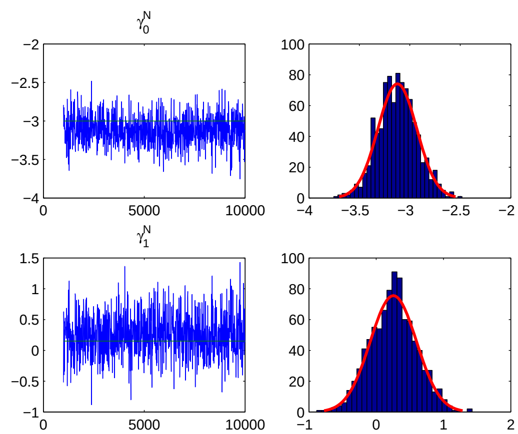

Figure 2.15: Chains and histograms for $\boldsymbol{\gamma}^{\boldsymbol{N_2}}$, third simulation

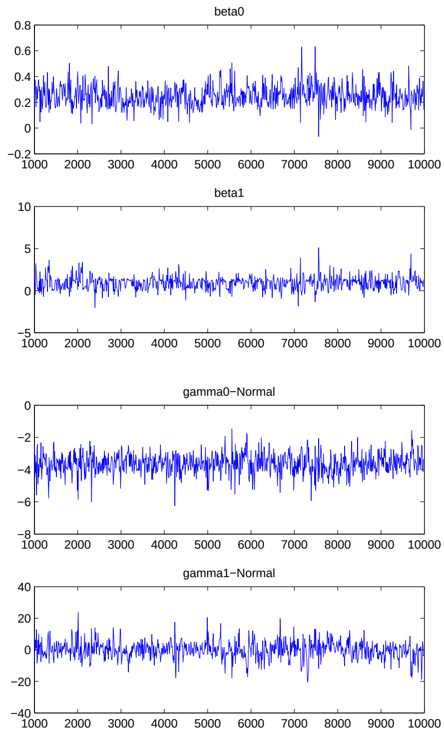

Figure 2.16: Chains for β and γ, from the normal distribution, home valuations

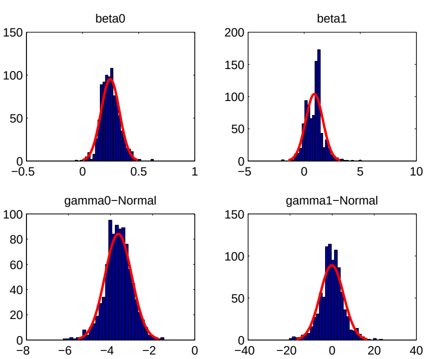

Figure 2.17: Histograms for the parameters of the normal distribution, home valuations

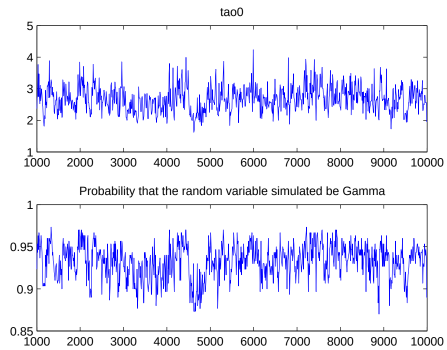

Figure 2.18: Chains for τ and probability of being normal, home valuations

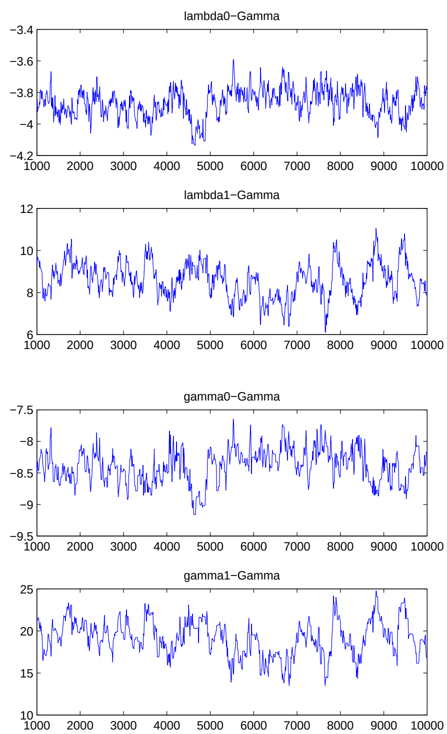

Figure 2.19: Chains for λ and γ, from the gamma distribution, home valuations

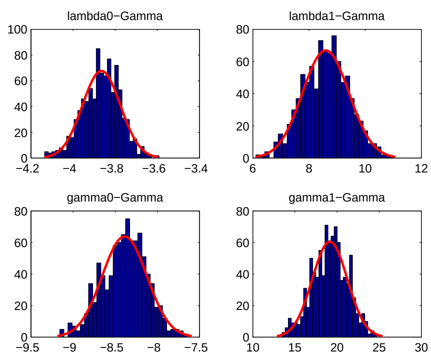

Figure 2.20: Histograms for the parameters of the Gamma distibution, home valuations

Chapter 3

HETEROSCEDASTIC WEIBULL-NORMAL MIXTURE MODELS: A BAYESIAN APPROACH

Summary: In this paper, we applied Bayesian methodology to obtain parameter estimates of the mixture of distributions belonging to the normal and biparametric weibull families, modelling the mean and the variance. Simulated studies and applications show the performance of the proposed models.

3.1 Introduction

The Weibull distribution has been used in many fields mainly describing the elapsed time until failure or death of an event. This modelled time recieves different names depending on the field of study, as for example, survival analysis, reliability analysis, duration analysis. It has also been used in extreme value theory, weather forecasting and communications systems, among others.

Sohn S.Y. et al. (2006) propose a random effects Weibull regression model to explain occupational lifetime of employees, modelling the scale parameter according to individual characteristics, estimating the parameters by maximum likelihood. Jalmar M.F. et al. (2008) model survival time with homoscedastic log-modified Weibull regression models, estimating the parameters by maximum likelihood. Jiang and Murthy (1995, 1998) model Failure-Data using a mixture of two biparametric weibull distributions with five parameters. The fifth is the parameter of the incidence probability.

In this chapter we propose Bayesian Methodology to estimate the parameters in mixture of distributions from the weibull and normal families, modelling both, the mean and the variance. The general algorithm that we propose is also useful for particular cases as weibull regressions, normal regressions, mixtures in which some of the distributions involved have constant means or constant variances.

Since the mean and the variance are modelled, a numerical method to obtain the usual parameters of the weibull distribution, α and δ, is needed. We use a combination of Newton and bisection methods, so as to allow the parameters to be in a determined range.

This chapter has five additional sections. In section 2 the formulation of the model is presented. In section 3 the Bayesian methodology employed is proposed. In section 4 the results of two simulations are included. In section

5 the proposed Bayesian methodology is applied to two practical examples. In section 6 general conclusions are given.

3.2 The Model

Let Y be a random variable of interest and y_i, $i = 1, \ldots, n$, n independent realizations of Y with probability density function given by

$$f(y; \mathbf{x}, \mathbf{w}, \boldsymbol{\theta}_j) = \sum_{j=1}^{k} a_j f_j(y; \mathbf{x}, \mathbf{w}, \boldsymbol{\theta}_j). \tag{3.1}$$

where $f_j(y; \mathbf{x}, \mathbf{w}, \boldsymbol{\theta}_j)$, $j = 1, ..., k$, are distributions from the normal and weibull families, and a_j are weights depending on a vector of covariates $\mathbf{x}$, $0 \leq a_j \leq 1$ and $a_1 + \cdots + a_k = 1$. Model 4.1 is called a "finite mixture distribution." The mixture weights a_j are the probabilities that the random variable Y_i follows the distribution of a "mixture component" f_j. In this case, we assume that the mean and variance of each of the components of the mixture are given by regression models, with respective vectors of covariates $\mathbf{x} = [1, x_1, \ldots, x_r]$ and $\mathbf{w} = [1, w_1, \ldots, w_l]$. A weibull regression of joint modelling of the mean and the variance is a particular case when $k = 1$.

The following notation is employed throughout. The vector of observations is $\mathbf{Y} = (Y_1, \ldots, Y_n)'$, the covariate matrices for the mean and variance regressions are respectively $\mathbf{X}$ and $\mathbf{W}$, with respective row vectors $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{il})$ and $\mathbf{w}_i = (1, w_{i1}, \ldots, w_{in})$, $i = 1, \ldots, n$. The general notation for the vectors of covariates in any of the distributions of the mixture is $\mathbf{x} = (1, x_1, \ldots, x_l)$ and $\mathbf{w} = (1, \ldots, w_r)$. $\boldsymbol{\theta}_j$ is the biparametric vector of parameters of distribution j in the mixture.

In order to obtain a simplification in the application of Bayesian methodology, latent Bernoulli variables, z_{ij}, $i = 1, ..., n$, $j = 1, ..., k$, with success probability given by (3.2), are introduced (see for example, Tanner and Wong, 1987; or Casela et al, 2002).

$$h_{ij} := \frac{a_j(\mathbf{x}_i) f_j(y_i; \mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\theta}_j)}{\sum_{t=1}^{k} a_t(\mathbf{x}_i) f_t(y_i; \mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\theta}_t)} \tag{3.2}$$

Thus, taking into account these hidden variables and using $\{\mathbf{x}_i\}_{i=1}^{n}$ to denote the set of vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, the conditional joint density function for $\mathbf{z} = (z_{11}, \ldots, z_{1k}, z_{21}, \ldots, z_{nk})$ given $\left(\mathbf{y}, \{\mathbf{x}_i\}_{i=1}^{n}, \{\mathbf{w}_i\}_{i=1}^{n}, \{\boldsymbol{\Theta}_j\}_{j=1}^{k}\right)$ is

$$\begin{aligned} f(\mathbf{z}|\mathbf{y}, \{\mathbf{x}_i\}_{i=1}^{n}, \{\mathbf{w}_i\}_{i=1}^{n}, \{\boldsymbol{\Theta}_j\}_{j=1}^{k}) &= \prod_{i=1}^{n} \prod_{j=1}^{k} h_{ij}^{z_{ij}} (1 - h_{ij})^{1-z_{ij}} \\ = \prod_{i=1}^{n} \prod_{j=1}^{k} & \frac{(a_j(\mathbf{x}_i) f_j(y_i; \mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\theta}_j))^{z_{ij}} \left(\sum_{l=1, l \neq j}^{k} a_l(\mathbf{x}_i) f_l(y_i; \mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\theta}_l)\right)^{1-z_{ij}}}{\sum_{l=1}^{k} a_l(\mathbf{x}_i) f_l(y_i; \mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\theta}_l)} \end{aligned} \tag{3.3}$$

and the conditional joint density function for $(\mathbf{y}, \mathbf{z})$ conditioned on

$(\{\mathbf{x}_i\}_{i=1}^n, \{\mathbf{w}_i\}_{i=1}^n, \{\mathbf{\Theta}_j\}_{j=1}^k)$, is given by,

$$
\begin{aligned}
&f(\mathbf{y},\mathbf{z}|\, \{\mathbf{x}_i\}_{i=1}^n, \{\mathbf{w}_i\}_{i=1}^n, \{\mathbf{\Theta}_j\}_{j=1}^k) = \\
&= \prod_{i=1}^n \prod_{j=1}^k \frac{(a_j(\mathbf{x}_i) f_j(y_i;\mathbf{x}_i,\mathbf{w}_i,\boldsymbol{\theta}_j))^{z_{ij}} \left(\sum_{l=1,l\neq j}^k a_l(\mathbf{x}_i) f_l(y_i;\mathbf{x}_i,\mathbf{w}_i,\boldsymbol{\theta}_l)\right)^{1-z_{ij}}}{\sum_{l=1}^k a_l(\mathbf{x}_i) f_l(y_i;\boldsymbol{\theta}_l)} \\
&\qquad\qquad \times \prod_{i=1}^n \sum_{j=1}^k a_j(\mathbf{x}_i) f_j(y_i;\mathbf{x}_i,\mathbf{w}_i,\boldsymbol{\theta}_j) \\
&= \prod_{i=1}^n \prod_{j=1}^k (a_j(\mathbf{x}_i) f_j(y_i;\mathbf{x}_i,\mathbf{w}_i,\boldsymbol{\theta}_j))^{z_{ij}} \left(\sum_{l=1,l\neq j}^k a_l(\mathbf{x}_i) f_l(y_i;\mathbf{x}_i,\mathbf{w}_i,\boldsymbol{\theta}_l)\right)^{1-z_{ij}}
\end{aligned}
\tag{3.4}
$$

We specifically assume the model where the mixture components and the weights are given by

$$
f_j(y) = \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{1}{2}\left(\frac{y-(\mathbf{x}\boldsymbol{\beta}_j)}{\sigma_j}\right)^2} \tag{3.5}
$$

$$
f_j(y) = \frac{\alpha_j}{\delta_j} y^{\alpha_j - 1} e^{-\frac{y^{\alpha_j}}{\delta_j}} I_{(0,\infty)}(y) \tag{3.6}
$$

$$
a_j = \frac{e^{\mathbf{x}\boldsymbol{\tau}_j}}{1+\sum_{l=1}^{l=k-1} e^{\mathbf{x}\boldsymbol{\tau}_l}} \tag{3.7}
$$

where $j = 1, ..., k$ in (3.5) and (3.6), meaning that any of the k distributions could be normal or weibull; $j = 1, \ldots, k-1$ in (3.7), and $a_k = 1 - \sum_{j=1}^{k-1} a_j = \frac{1}{1+\sum_{l=1}^{l=k-1} e^{x\boldsymbol{\tau}_l}}$.

The model for the variance of the normal and weibull distributions is $\sigma^2(\mathbf{w}) = \exp(\mathbf{w}\boldsymbol{\gamma}^d)$, where $d = N$ or $d = w$ depending on the distribution being normal or weibull. The model for the mean of the normal distribution is $\mu = \mathbf{x}\boldsymbol{\beta}$ and for the weibull distribution is $\mu(\mathbf{x}) = \exp(\mathbf{x}\boldsymbol{\lambda})$. The mean and variance of the weibull distribution as functions of the original parameters α, δ, are

$$\mu = E(Y) = \delta^{\frac{1}{\alpha}}\Gamma\left(\alpha^{-1}+1\right) \tag{3.8}$$

$$\sigma^2 = Var(Y) = \delta^{\frac{2}{\alpha}}\left\{\Gamma\left(2\alpha^{-1}+1\right) - \left[\Gamma\left(\alpha^{-1}+1\right)\right]^2\right\}. \tag{3.9}$$

If μ and σ^2 are modelled, one can solve α and δ from the equations (3.10) and (3.11) using Newton numerical method.

$$\frac{\mu^2}{\sigma^2} = \frac{\left[\Gamma\left(\alpha^{-1}+1\right)\right]^2}{\left\{\Gamma\left(2\alpha^{-1}+1\right) - \left[\Gamma\left(\alpha^{-1}+1\right)\right]^2\right\}} \tag{3.10}$$

$$\delta^{\frac{1}{\alpha}} = \frac{\mu}{\Gamma\left(\alpha^{-1}+1\right)} \tag{3.11}$$

3.3 Bayesian Methodology

This section develops the Bayesian methodology employed to fit the mixture model with distributions and weights given by (3.5), (3.6) and (3.7), in which means and variances are explained by regressors.

3.3.1 Prior distributions

For the prior distributions of the parameters we assume normal distributions with diagonal variance-covariance matrices and high variance values, so as to introduce the least amount of possible information. For the sake of simplicity, we suppose independence among the parameters a priori. These a priori distributions are given by

$$
\begin{aligned}
\boldsymbol{\beta}_{i_N} &\sim N(\mathbf{b}_{i_N}, B_{i_N}^{-1}), \\
\boldsymbol{\gamma}_{i_d}^{d} &\sim N(\mathbf{g}_{i_d}^{d}, (G_{i_d}^{d})^{-1}),\ d = N, w \\
\boldsymbol{\lambda}_{i_w} &\sim N(\mathbf{l}_{i_w}, L_{i_w}^{-1}) \\
\boldsymbol{\tau}_j &\sim N(\mathbf{t}_j, T_j^{-1}), j = 1, ..., k-1
\end{aligned}
\tag{3.12}
$$

where $i_d \in \{1, ..., m_d\}$, m_N is the number of normal distributions in the mixture, m_w is the number of weibull distributions in the mixture, $m_N + m_w = k$. The vector of parameters for the variance of the i_N-th nomal destribution is $\boldsymbol{\gamma}_{i_N}^{N}$ and for the i_w-th weibull distribution is $\boldsymbol{\gamma}_{i_w}^{w}$. The vector of parameters for the mean of the i_w-th weibull distribution is $\boldsymbol{\lambda}_{i_w}$.

From (3.4) and (3.12) and using Bayes' theorem, the posterior distribution is given by

$$
\pi(\boldsymbol{\Theta}|\mathbf{y}, \mathbf{z}, \{\mathbf{x}_i\}_{i=1}^{n}, \{\mathbf{w}_i\}_{i=1}^{n}) \propto P(\boldsymbol{\theta}) L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}, \{\mathbf{x}_i\}_{i=1}^{n}, \{\mathbf{w}_i\}_{i=1}^{n})
$$

where $\boldsymbol{\theta}$ is the vector of all the parameters involved, that is,

$$
\boldsymbol{\theta} = \left(\{\boldsymbol{\beta}'_{i_N}\}_{i_N=1}^{m_N}, \{\boldsymbol{\gamma}'_{i_N}\}_{i_N=1}^{m_N}, \left\{\boldsymbol{\lambda}'_{i_g}\right\}_{i_g=1}^{m_g}, \left\{\boldsymbol{\gamma}'_{i_g}\right\}_{i_g=1}^{m_g}, \{\boldsymbol{\tau}'_j\}_{j=1}^{k} \right)'
$$

and $P(\boldsymbol{\theta})$ is the joint prior density function. The samples for the parameters will be taken from their conditional posterior distributions defined below, using Gibbs sampler and Metropolis -Hastings algorithm.

3.3.2 Conditional posterior distributions of the parameters

The conditional posterior for $\boldsymbol{\beta} := \boldsymbol{\beta}_{i_N}$, the vector of parameters of the mean in any of the m_d normal distributions in the mixture is given by

$$\pi\left(\boldsymbol{\beta}\left|\mathbf{y},\mathbf{z},\{\mathbf{x}_i\}_{i=1}^n,\{\mathbf{w}_i\}_{i=1}^n,\boldsymbol{\gamma}^N\right.\right)\propto$$
$$\frac{|B|^{\frac{1}{2}}}{\left(\sqrt{2\pi}\right)^2}\exp\left(-\frac{1}{2}\left(\boldsymbol{\beta}-\mathbf{b}\right)'B\left(\boldsymbol{\beta}-\mathbf{b}\right)\right)\exp\left(-\frac{1}{2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)'\Sigma_z^{-1}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)\right) \tag{3.13}$$

where $\Sigma_z^{-1}=\operatorname{diag}\left(z_{1i_N}/\sigma_1^2,\ldots,z_{ni_N}/\sigma_n^2\right)$, $\sigma_i^2=\exp\left(\mathbf{W}\boldsymbol{\gamma}^N\right)_i$ and $z_{ji_N}=1$ if y_j is an observation from the i_Nth normal distribution. So $\boldsymbol{\beta}\sim N(\boldsymbol{\mu},\Sigma)$ with $\Sigma^{-1}=B+\mathbf{X}'\Sigma_z^{-1}\mathbf{X}$ and $\Sigma^{-1}\boldsymbol{\mu}=B\mathbf{b}+\mathbf{X}'\Sigma_z^{-1}\mathbf{y}$

The conditional posterior for $\boldsymbol{\lambda}_{i_w}$, the vector of parameters of the mean of the i_wth weibull distribution in the mixture is given by

$$\pi\left(\boldsymbol{\lambda}_{i_w}|\mathbf{y},\mathbf{z},\{\mathbf{x}_i\}_{i=1}^n,\{\mathbf{w}_i\}_{i=1}^n,\boldsymbol{\gamma}_{i_w}\right)\propto$$
$$\frac{|L_{i_w}|^{\frac{1}{2}}}{(\sqrt{2\pi})^2}\exp\left(-\frac{1}{2}(\boldsymbol{\lambda}_{i_w}-\mathbf{l}_{i_w})'L_{i_w}(\boldsymbol{\lambda}_{i_w}-\mathbf{l}_{i_w})\right)\prod_{j\in J}\frac{\alpha_j}{\delta_j}y_j^{\alpha_j-1}e^{-\frac{y_j^{\alpha_j}}{\delta_j}}I_{(0,\infty)}(y_j) \tag{3.14}$$

where $J=\{j\,|j=1,...,n,z_{ji_w}=1\}$, $\alpha_j=\alpha_j(\mu_j,\sigma_j^2)$, $\delta_j=\delta_j(\mu_j,\sigma_j^2)$, $\mu_j=\exp\left(\mathbf{X}\boldsymbol{\lambda}_{i_w}\right)_j$, $\sigma_j^2=\exp(\mathbf{W}\boldsymbol{\gamma}_{i_w})_j$.

The conditional posterior for $\boldsymbol{\tau}_j$, the vector of parameters of the weight

a_j, $j = 1, \ldots, k-1$, is given by

$$\pi\left(\boldsymbol{\tau}_j|\mathbf{z},\{\mathbf{x}_i\}_{i=1}^{n},\{\boldsymbol{\tau}_l\}_{l=1,l\neq j}^{k-1},\right) \propto \frac{|T_j|^{\frac{1}{2}}}{(\sqrt{2\pi})^2}\exp\left(-\frac{1}{2}\left(\boldsymbol{\tau}_j-\mathbf{t}_j\right)' T_j\left(\boldsymbol{\tau}_j-\mathbf{t}_j\right)\right)\frac{\prod_{i\in I}\exp(\mathbf{X}\boldsymbol{\tau}_j)_i}{\prod_{i=1}^{n}\left(1+\sum_{l=1}^{k-1}\exp(\mathbf{X}\boldsymbol{\tau}_l)_i\right)} \tag{3.15}$$

where $I = \{i\,|i = 1, ..., n, z_{ij} = 1\}$.

The conditional posterior for $\boldsymbol{\gamma}_{i_N}^{N}$, the vector of parameters of the variance in the i_N normal distributions in the mixture, is given by

$$\pi\left(\boldsymbol{\gamma}_{i_N}\left|\mathbf{y},\mathbf{z},\{\mathbf{x}_i\}_{i=1}^{n},\{\mathbf{w}_i\}_{i=1}^{n},\boldsymbol{\beta}_{i_N}\right.\right) \propto \frac{|G_{i_N}|^{\frac{1}{2}}}{\left(\sqrt{2\pi}\right)^2}\exp\left(-\frac{1}{2}\left(\boldsymbol{\gamma}_{i_N}-\mathbf{g}_{i_N}\right)' G_{i_N}\left(\boldsymbol{\gamma}_{i_N}-\mathbf{g}_{i_N}\right)\right)\prod_{j\in J}\frac{\exp\left(-\frac{1}{2}\frac{\left(y_j-(\mathbf{X}\boldsymbol{\beta}_{i_N})_j\right)^2}{\sigma_j^2}\right)}{\left(\sigma_j^2\right)^{\frac{1}{2}}} \tag{3.16}$$

where $J = \{j\,|j = 1, \ldots, n, z_{ji_N} = 1\}$, $\sigma_j^2 = \exp\left(\mathbf{W}\gamma_{i_N}\right)_j$.

The conditional posterior for $\boldsymbol{\gamma}_{i_w}$ the vector of parameters of the variance of the i_wth weibull distribution in the mixture, is given by

$$\pi\left(\boldsymbol{\gamma}_{i_w}|\mathbf{y},\mathbf{z},\{\mathbf{x}_i\}_{i=1}^{n},\{\mathbf{w}_i\}_{i=1}^{n},\boldsymbol{\lambda}_{i_w}\right) \propto \frac{|G_{i_w}|^{\frac{1}{2}}}{\left(\sqrt{2\pi}\right)^2}\exp\left(-\frac{1}{2}\left(\boldsymbol{\gamma}_{i_w}-\mathbf{g}_{i_w}\right)' G_{i_w}\left(\boldsymbol{\gamma}_{i_w}-\mathbf{g}_{i_w}\right)\right)\prod_{j\in J}\frac{\alpha_j}{\delta_j}y_j^{\alpha_j-1}e^{-y_j^{\frac{\alpha_j}{\delta_j}}}I_{(0,\infty)}(y_j) \tag{3.17}$$

where $J = \{j\,|j = 1, \ldots, n, z_{ji_w} = 1\}$, $\alpha_j = \alpha_j(\mu_j, \sigma_j^2)$, $\delta_j = \delta_j(\mu_j, \sigma_j^2)$, $\mu_j = \exp(\mathbf{X}\boldsymbol{\lambda}_{i_w})_j$, $\sigma_j^2 = \exp(\mathbf{W}\boldsymbol{\gamma}_{i_w})_j$.

3.3.3 Bayesian proposed methodology

Observe that $\pi\left(\boldsymbol{\beta}\left|\mathbf{y},\mathbf{z},\{\mathbf{x}_i\}_{i=1}^n,\{\mathbf{w}_i\}_{i=1}^n,\boldsymbol{\gamma}^N\right.\right)$ is a density which can be sampled from, so the Gibbs sampler will be used for $\boldsymbol{\beta}$. For the other parameters, the Metropolis Hastings algorithm will be proposed.

Based on the ideas in Cepeda and Gamerman (2001, 2005), the transition kernel q for $\boldsymbol{\lambda}$, $\boldsymbol{\gamma}^N$,$\boldsymbol{\gamma}^W$ is a normal density which is obtained by combining the prior distribution of the parameter vector with the likelihood function of the working obsevations variables whose joint density is assumed to be multivariate normal. To be precise, if $\mu = g(\mathbf{x}\boldsymbol{\lambda})$, $\sigma^2 = g(\mathbf{w}\boldsymbol{\gamma})$, the working observation variable will be defined as the linear approximation of $g^{-1}(t)$ around the mean, in the first case, and around the variance in the other, choosing a random variable t, for each case, in such a way that $E(t-\mu)=0$ in the first case, and $E(t-\sigma^2) = 0$ in the other. In general, if δ is the parameter to be modelled, for example μ or σ^2, the working observation variables are defined as follows:

$$\tilde{y} := g^{-1}(\delta) + (g^{-1})'(\delta)\,(t-\delta)$$

Thus $E(\tilde{y}) = g^{-1}(\delta)$ and $\mathrm{Var}(\tilde{y}) = \left((g^{-1})'\right)^2(\delta)\mathrm{Var}(t)$. With this in mind, we proceed to define the transition kernel q in the Metropolis-Hastings algorithm for λ, γ^N and γ^g.

1. Definition of the transition kernel for the vector of parameters, γ^N, of the variance in any of the normal dictributions in the mixture.

   The following analysis is based on the fact that the observations Y_i involved in the full conditional distribution for γ^N are normal. To define the corresponding working observation variable $\hat{Y}_i$, that for the general case will be denoted $\hat{Y}$, we set $t = (Y - \mathbf{x}\boldsymbol{\beta})^2$ for which $E(t) = \sigma^2$ and $\text{Var}(t) = 2\sigma^4$, where σ^2=exp($\mathbf{w}\boldsymbol{\gamma}$). Then

$$\begin{aligned} g(t) &\approx g(\sigma^2) + g'(\sigma^2)(t - \sigma^2) \\ &= \mathbf{w}\boldsymbol{\gamma} + \frac{1}{\sigma^2}(t - \sigma^2) \\ &= \mathbf{w}\boldsymbol{\gamma} + \frac{(Y - \mathbf{x}\boldsymbol{\beta})^2}{\sigma^2} - 1 \end{aligned}$$

Defining the working observation $\tilde{Y}$ as $\tilde{Y} := \mathbf{w}\boldsymbol{\gamma} + \frac{(Y - \mathbf{x}\boldsymbol{\beta})^2}{\sigma^2} - 1$, we have $E(\tilde{Y}) = \mathbf{w}\boldsymbol{\gamma}$, and $\text{Var}\tilde{Y} = 2$. Assuming that the distribution of $\tilde{Y}$ is a normal distribution $N(\mathbf{w}\boldsymbol{\gamma}, 2)$, a multivariate normal $N(\mathbf{W}\boldsymbol{\gamma}, 2I)$ is the joint distribution of $\tilde{\mathbf{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)'$, where I is the identity $n \times n$ matrix. Combining the prior for $\boldsymbol{\gamma}$ with this multivariate normal distribution, a transition kernel $q = \tilde{\pi}(\boldsymbol{\gamma})$ is given by

$$\tilde{\pi}(\gamma) \propto \\ \frac{|G|^{\frac{1}{2}}}{(\sqrt{2\pi})^2}\left(-\frac{1}{2}(\boldsymbol{\gamma} - \mathbf{g})')G(\boldsymbol{\gamma} - \mathbf{g})\right)\frac{|2I|^{-\frac{1}{2}}}{(\sqrt{2\pi})^n}\text{exp}\left(-\frac{1}{2}(\tilde{\mathbf{y}} - \mathbf{W}\boldsymbol{\gamma})'(2I)^{-1}(\tilde{\mathbf{y}} - \mathbf{W}\boldsymbol{\gamma})\right) \tag{3.18}$$

That is, the proposed kernel for $\boldsymbol{\gamma}$ is a multivariate normal density function $N(\boldsymbol{\mu}, \Sigma)$, where $\Sigma^{-1} = G + \frac{1}{2}\mathbf{W}'\mathbf{W}$ and $\Sigma^{-1}\boldsymbol{\mu} = G\mathbf{g} + \frac{1}{2}\mathbf{W}'\tilde{\mathbf{y}}$.

2. Definition of the transition kernel of the vector of parameters, $\boldsymbol{\lambda}$, of the mean of any of the m_w weibull distributions in the mixture.

   The following analysis is based on the fact that the observations Y_i involved in the conditional distribution for $\boldsymbol{\lambda}$ have a weibull distribution with parameters (α, δ). Setting first $t = Y$, we have $E(t) = \mu$ and $\text{Var}(t) = \sigma^2$, where σ^2=exp($\mathbf{w}\boldsymbol{\gamma}$), μ = exp($\mathbf{x}\boldsymbol{\lambda}$), and their relation with the original parameters of the weibull distribution is

$$\mu = \delta^{\frac{1}{\alpha}}\Gamma\left(\alpha^{-1}+1\right),$$

$$\sigma^2 = \delta^{\frac{2}{\alpha}}\left\{\Gamma\left(2\alpha^{-1}+1\right) - \left[\Gamma\left(\alpha^{-1}+1\right)\right]^2\right\}.$$

We now define the working observation as

$$\tilde{Y} := \mathbf{x}\boldsymbol{\lambda} + \left(\frac{Y - \exp\left(\mathbf{x}\boldsymbol{\lambda}\right)}{\exp\left(\mathbf{x}\boldsymbol{\lambda}\right)}\right),$$

for which $E(\tilde{Y}) = \mathbf{x}\boldsymbol{\lambda}$ and $\text{Var}(\tilde{Y}) = \frac{\sigma^2}{\mu_g^2} = \frac{\exp(\mathbf{w}\boldsymbol{\gamma})}{\exp(2(\mathbf{x}\boldsymbol{\lambda}))}$. Assuming that $\tilde{\mathbf{Y}} = \left(\tilde{Y}_1, \ldots, \tilde{Y}_n\right)$ has a multivariate normal distribution $N(\mathbf{X}\boldsymbol{\lambda}, A)$, where $A = \text{diag}\left(\frac{\exp(\mathbf{w}\boldsymbol{\gamma})_i}{\exp(2(\mathbf{x}\boldsymbol{\lambda})_i)}\right)$, and combining it with the prior for $\boldsymbol{\lambda}$, a posterior distribution $q = \tilde{\pi}(\boldsymbol{\lambda})$ is given by

$$\tilde{\pi}(\boldsymbol{\lambda}) \propto$$
$$\frac{|L|^{\frac{1}{2}}}{(\sqrt{2\pi})^2}\left(-\frac{1}{2}(\boldsymbol{\lambda} - \mathbf{l})')L(\boldsymbol{\lambda} - \mathbf{l})\right)\frac{|A|^{-\frac{1}{2}}}{(\sqrt{2\pi})^n}\exp\left(-\frac{1}{2}(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\lambda})'(A)^{-1}(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\lambda})\right). \tag{3.19}$$

Thus, a proposed kernel for $\boldsymbol{\lambda}$ is a multivariate normal distribution $N(\boldsymbol{\mu}, \Sigma)$ where $\Sigma^{-1} = L + \mathbf{X}'A^{-1}\mathbf{X}$ and $\Sigma^{-1}\boldsymbol{\mu} = L\mathbf{l} + \mathbf{X}'A^{-1}\tilde{\mathbf{y}}$.

3. Definition of the transition kernel of the parameters, $\boldsymbol{\gamma}^w$, of the variance in any of the weibull distributions in the mixture. Before defining the working observation variable we set $t = (Y - \mathbf{x}\boldsymbol{\lambda})^2$, then $\text{E}(t) = \sigma^2$ and

$$\mathrm{Var}(t) = \delta^{4/\alpha}\Gamma(\frac{4}{\alpha}+1) - 4\delta^{3/\alpha}\Gamma(\frac{3}{\alpha}+1)\delta^{1/\alpha}\Gamma(\frac{1}{\alpha}+1) + 6\delta^{2/\alpha}\Gamma(\frac{2}{\alpha}+1)\delta^{2/\alpha}\Gamma^2(\frac{1}{\alpha}+1)$$
$$-4\delta^{1/\alpha}\Gamma(\frac{1}{\alpha}+1)\delta^{3/\alpha}\Gamma^3(\frac{1}{\alpha}+1) + \delta^{4/\alpha}\Gamma^4(\frac{1}{\alpha}+1) - \sigma^4,$$

where $\sigma^2 = \exp(\mathbf{w}\boldsymbol{\gamma})$, $\mu = \exp(\mathbf{x}\boldsymbol{\lambda})$, and their relation with the weibull parameters is

$$\mu = \delta^{\frac{1}{\alpha}}\Gamma\left(\alpha^{-1}+1\right),$$
$$\sigma^2 = \delta^{\frac{2}{\alpha}}\left\{\Gamma\left(2\alpha^{-1}+1\right) - \left[\Gamma\left(\alpha^{-1}+1\right)\right]^2\right\}.$$

The working observation variable is defined by

$$\tilde{Y} = \mathbf{w}\boldsymbol{\gamma} + \left(\frac{t}{\sigma^2} - 1\right),$$

for which $E(\tilde{Y}) = \mathbf{w}\boldsymbol{\gamma}$ and $\mathrm{Var}(\tilde{Y}) = \frac{1}{\sigma^4}\mathrm{Var}(t)$. Assuming that $\tilde{\mathbf{Y}} = (\tilde{Y}_1 \ldots, \tilde{Y}_n)'$ has a multivariate normal distribution, $N(\mathbf{W}\boldsymbol{\gamma}, A)$, where $A = \mathrm{diag}(\mathrm{Var}(\tilde{Y}_i))$, and combining it with the prior for $\boldsymbol{\gamma}$, a posterior distribution $q = \tilde{\pi}(\boldsymbol{\gamma})$ is given by

$$\tilde{\pi}(\boldsymbol{\gamma}) \propto$$
$$\frac{|G|^{\frac{1}{2}}}{(\sqrt{2\pi})^2}\left(-\frac{1}{2}(\boldsymbol{\gamma}-\mathbf{g})')G(\boldsymbol{\gamma}-\mathbf{g})\right)\frac{|A|^{-\frac{1}{2}}}{(\sqrt{2\pi})^n}\exp\left(-\frac{1}{2}(\tilde{\mathbf{y}}-\mathbf{W}\boldsymbol{\gamma})'(A)^{-1}(\tilde{\mathbf{y}}-\mathbf{W}\boldsymbol{\gamma})\right) \tag{3.20}$$

That is, a proposed kernel for $\boldsymbol{\gamma}$ is a multivariate normal density function, $N(\boldsymbol{\mu}, \Sigma)$, where $\Sigma^{-1} = G + \mathbf{W}'A^{-1}\mathbf{W}$, and $\Sigma^{-1}\boldsymbol{\mu} = G\mathbf{g} + \mathbf{W}'A^{-1}\tilde{\mathbf{y}}$.

4. The parameters $\boldsymbol{\tau}_j$ of the weights a_j, $j = 1, \ldots, k-1$, are drawn from a random walk.

3.4 Simulations

The results of two simulations are shown in this section. In both simulations the means and the variances of the mixture of normal and weibull distributions were modelled, and the corresponding parameters were estimated using the proposed bayesian methodology. Independent normal flat priors distributions were assigned to the parameters estimated.

3.4.1 First simulated study

The model is a mixture $f(y) = a_1 f_N(y; x, w, \boldsymbol{\beta}, \boldsymbol{\gamma}_N) + a_2 f_w(y; x, w, \boldsymbol{\beta}, \boldsymbol{\gamma}_w)$ where f_N is as in (3.5), f_w is as in (3.6) and a_1 as in (3.7). The model for the variance of the distributions involved is $\sigma^2(w) = \exp(\gamma_0^d + \gamma_1^d w)$ where $d = N, w$, depending on the distribution being normal, or weibull. The model for the mean of the normal distribution in the simulation is $\mu^N(x) = \beta_0 + \beta_1 x$, and $\mu^d = \exp(\lambda_0^w + \lambda_1^w x)$, for the Weibull distribution. 300 observations were generated from the model, using the true parameter values. The incidence probability is $a_1 = e^\tau/(1 + e^\tau)$, not depending on any covariate . To avoid large values entering the argument of the Gamma function, the covarites X and W were drawn from a uniform distribution $U(0, 5)$, whith no lack of generality since any scaling changes the values of the parameters but the observations continue having the same distributions. The values of Y where generated according to this mixture model, that is, a random value u_i from the uniform distribution in $(0, 1)$ is generated for the corresponding y_i. If $u \leq a_1$, a random value y_i from the normal distribution with the true paremeter values is generated, otherwise the value is generated from the weibull distribution. It is expected that more observations from the distribution whose incidence probability is a_1 are generated if $a_1 > 1/2$, that is, if $\tau > 0$. To apply Bayesian methodology we assigned independent normal prior distributions $N(0, 10^k)$ for all the parameters in the model. We take $k \geq 2$ to impose large prior variances, but, as we have already checked in our analysis, increasing this value to larger orders of magnitude made no effective difference in the estimation process. To form the samples of the estimated parameters, 7.000 observations of the posterior distributions of the parameters were generated, choosing every tenth of the current parameter estimates to form the chains. The convergence is slower since numerical methods have to be used to solve for the parameters of the weibull distibution. The posterior parameter estimates are shown in table 3.1 and graphs from the chains and histograms are in figures 3.1 to 3.3. As can be seen from the table, the estimates are very close to the true values and 95% credible intervals contain the true parameter values for all the parameters involved in the simulation. Jarque-Bera normality tests for each of the chains of the parameters did not reject the null hypothesis of normality. In the graphs of the chains, the horizontal line denotes the true parameter value. In this simulation $\tau = -0.3$ so, $a_1 = 0.43$, that is, 43% of the values of Y should be from the normal distribution, wich can be seen in figure (3.2) that shows the chain of $\hat{\tau}$ and the chain of the probabilities of being normal, which mainly counts in each observation of the chain the percentage of y values that go to the normal distribution.

	β_0	β_1	γ_0^N	γ_1^N	τ_0	λ_0	λ_1	γ_0^w	γ_1^w
t.v	0.2	0.1	0.4	0.02	-0.3	1.5	0.1	0.1	0.01
b.e	0.23	0.07	0.45	0.08	-0.25	1.50	0.1	-0.078	0.12
(se)	(0.31)	(0.1)	(0.31)	(0.09)	(0.15)	(0.04)	(0.009)	(0.23)	(0.09)

Table 3.1: Parameter estimates of the Normal-Weibull model

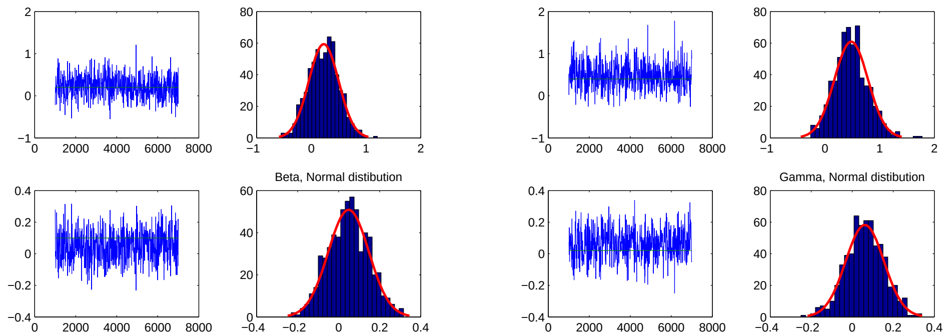

Figure 3.1: Chains and histograms for γ and β from the normal distribution, Normal-Weibull simulation

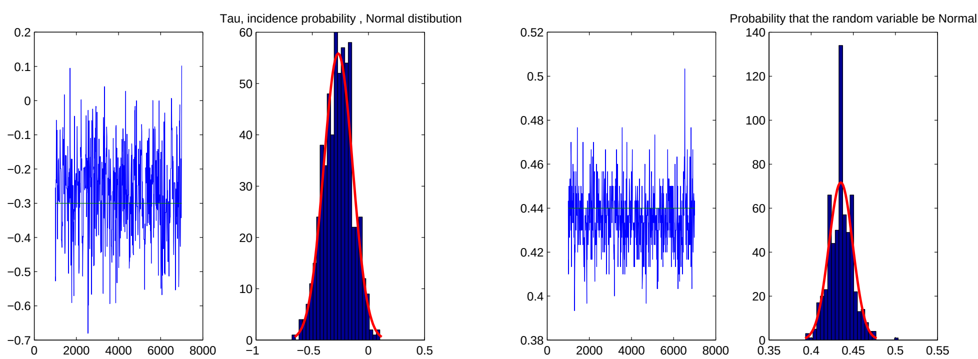

Figure 3.2: Chain for τ and probability of being normal, Normal-Weibull simulation

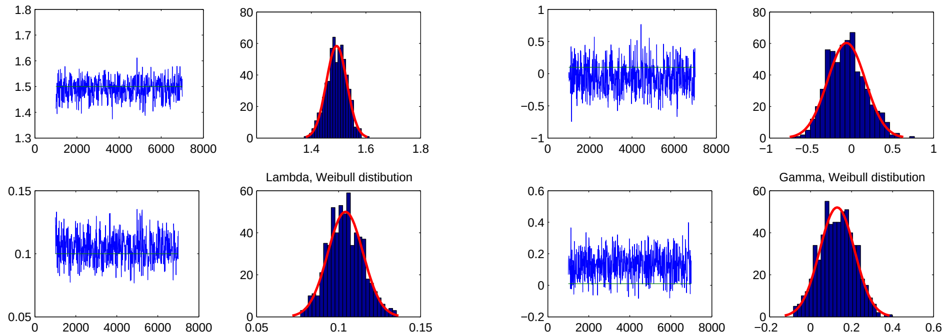

Figure 3.3: Chains and histograms for λ and γ from the Weibull distribution, Normal Weibull simulation

3.4.2 Second simulated study

To show the generality of the algorithm, in this simulation we estimate the parameters of the mixture of a normal and a weibull distribution with means, variances and incidence probability given by

$$\begin{aligned}
\mu_N(x_1,x_2) &= \beta_0+\beta_1 x_1+\beta_2 x_2 \qquad (3.21)\\
\mu_g(x_1,x_2) &= \lambda_0+\lambda_1 x_1+\lambda_2 x_2\\
\sigma^2_N(x_1,x_2) &= \exp\left(\gamma^N_0+\gamma^N_1 w_1+\gamma^N_2 w_2\right)\\
\sigma^2_g(x_1,x_2) &= \exp\left(\gamma^g_0+\gamma^g_1 w_1+\gamma^g_2 w_2\right)\\
a_1(x_1,x_2) &= \frac{\exp\left(\tau_0+\tau_1 x_1+\tau_2 x_2\right)}{1+\exp\left(\tau_0+\tau_1 x_1+\tau_2 x_2\right)}
\end{aligned}$$

100 observations from each of x_1, x_2, w_1 and w_2 were independently generated from the uniform distribution in the interval $(0,5)$, the vector $\mathbf{a}_1$ of the incidence probabilities was obtained according to the true τ parameter values. A vector, $\mathbf{u}$, the same size as $\mathbf{a}_1$, of independent observations from the uniform distribution in the interval (0,1) was generated, if $u_i \leq a_{i1}$, a

corresponding value y_i is generated from the normal distribution, otherwise, the y_i value is generated from the weibull distribution. The chains have 7000 observations with a burning period of 1000 observations, and every tenth observation is chosen. The results are shown in table(3.2) below and chains and histograms from figure (3.4.2) to figure (3.4.2). In this simulation $\boldsymbol{\tau} = [0, 1, -0.8]'$ and $\hat{\boldsymbol{\tau}} = \left[\underset{(0.6781)}{0.2779}, \underset{(0.5478)}{2.1986}, \underset{(0.4626)}{-1.9548}\right]$. In the first part of the simulation, the algorithm generated 62% of the y_i values from the normal distribution. The estimations of τ_1 and τ_2 were not good, even 95% credible intervals for these parameters did not contain the true parameter values, but the signs are correct and the estimation of the probability of being normal was good as can be seen in figure (3.4.2). The Jarque-Bera normality test did not reject the null hypothesis of normality for all the parameters but β_2 and λ_2. 95% credible intervals for the parameters contain the true parameter values except for τ_1 and τ_2.

	β_0	β_1	β_2	γ_0^N	γ_1^N	γ_2^N
t.v.	-1	0.1	-0.4	-0.1	-0.5	0.1
$\underset{(s.e.)}{b.e.}$	$\underset{(0.1633)}{-1.1023}$	$\underset{(0.0540)}{0.1161}$	$\underset{(0.0496)}{-0.3554}$	$\underset{(0.6356)}{-0.08944}$	$\underset{(0.1756)}{-0.6161}$	$\underset{(0.1502)}{0.1963}$
	λ_0	λ_1	λ_2	γ_0^g	γ_1^g	γ_2^g
t.v	2	-0.3	0.01	0.3	0.02	-0.1
$\underset{(s.e.)}{b.e.}$	$\underset{(0.0802)}{1.9089}$	$\underset{(0.0462)}{-0.3196}$	$\underset{(0.0213)}{0.0323}$	$\underset{(0.7002)}{-0.0706}$	$\underset{(0.1650)}{-0.0701}$	$\underset{(0.1778)}{0.1054}$

Table 3.2: Results of the second simulation

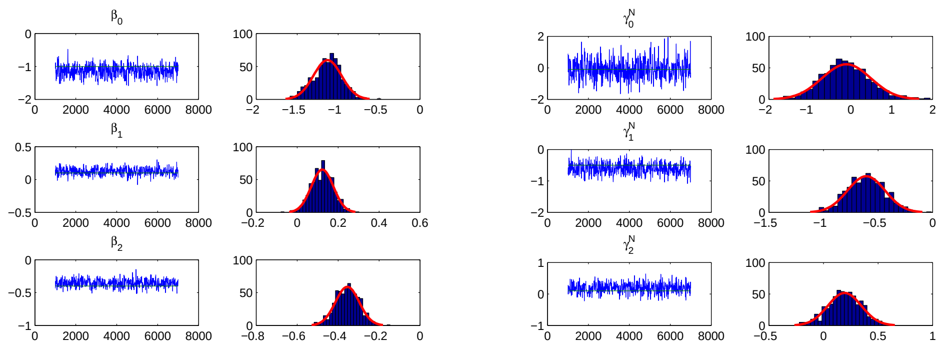

Figure 3.4: Chains and histograms for the parameter estimates of the mean and variance of the normal distribution, second simulation

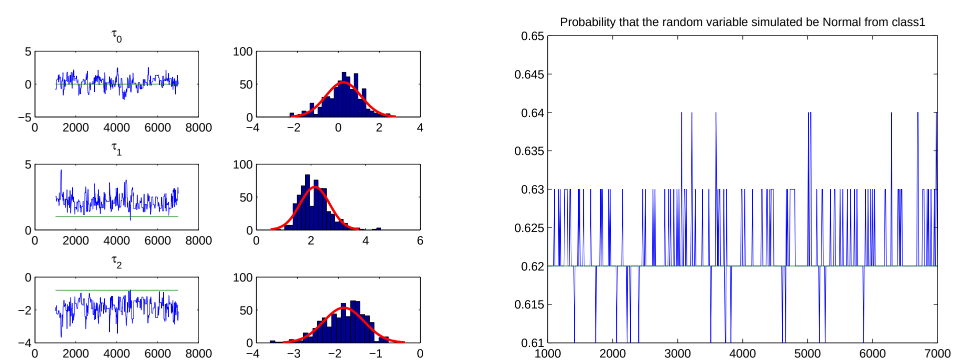

Figure 3.5: Chains for $\boldsymbol{\tau}$ and the probability of being normal, second simulation

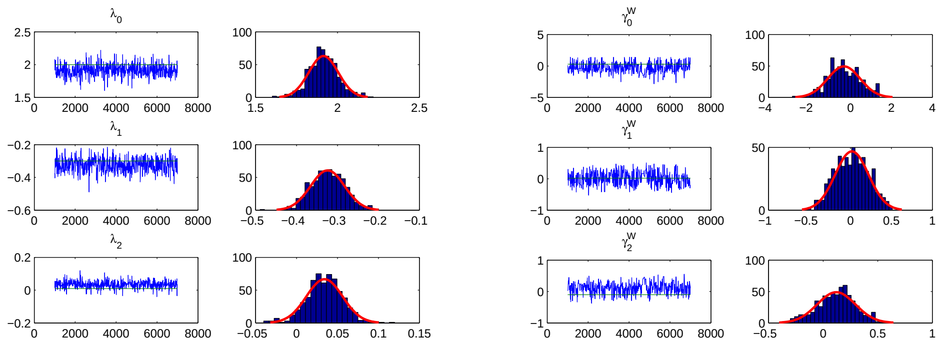

Figure 3.6: Chains and histograms for the parameter estimates of the mean and variance of the weibull distribution, second simulation

3.5 Applications

3.5.1 Example 1

The data in this application represent the survival time result of two distinct cancer treatments applied to a group of men and women during a time period. The explained variable Y represents the survival time, and the covariates are X_1 with two values 1 and 2 of cancer treatment, and X_2 with two values, 1 for female and 0 for male. This data set has 40 observations. The model is a weibull regression with mean and variance given by

$$\ln(\mu) = \lambda_0 + \lambda_1 X_1 + \lambda_2 X_2 \tag{3.22}$$

$$\ln(\sigma^2) = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 \tag{3.23}$$

with two further variations of the variance model, one with $\gamma_1 = 0$ and the other with $\gamma_2 = 0$. The results of the estimates and BIC values are given in table (3.3) below

	$\hat{\lambda}_0$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	BIC
Model 1	-0.0724 (0.0788)	-0.1194 (0.0392)	-0.2704 (0.0434)	-3.6123 (0.6100)	-0.4975 (0.4022)	-0.6362 (0.3706)	69.05
Model 2	-0.0380 (0.0757)	-0.1490 (0.0400)	-0.2742 (0.0561)	-4.2829 (0.3579)		-0.6990 (0.4514)	93.97
Model 3	-0.0822 (0.0614)	-0.1080 (0.0361)	-0.2917 (0.0436)	-3.7384 (0.6377)	-0.5828 (0.3871)		22.2256

Table 3.3: Weibull regressions for cancer data

For the samples of the parameter estimates, 10.000 observations were generated from the posterior distributions, choosing every tenth observation to form the chains, which are shown after a burning period of 1000 observations. 95% confidence intervals for the parameters in the first model showed that λ_1, λ_2 and γ_2 are significant, but the null hypothesis $\gamma_1 = 0$ was not rejected, so a second model with $\gamma_1 = 0$ was considered. The BIC value in this second model is bigger than the one in the first, and a 95% credible interval for γ_2 did not reject the null hypothesis $\gamma_2 = 0$. A third model with $\gamma_2 = 0$ was considered, giving a BIC value smaller than the former ones, and a 95% credible interval $[-1.2221, 0.1874]$ slightly not rejecting the hypothesis $\gamma_1 = 0$. A fourth model with constant variance was considered returning a BIC value of BIC=51.2. The results obtained in these regressions are consistent with the non parametric survival analysis which shows that the time of survival is greater for the patients following cancer treatment 1 and is also slighly more variable.

Figures (3.5.1) shows the chains and histograms of the posterior parameter samples of the second model. The Jarque-Bera test did not reject the null hypothesis of normality in any of the parameters in the seconnd model.

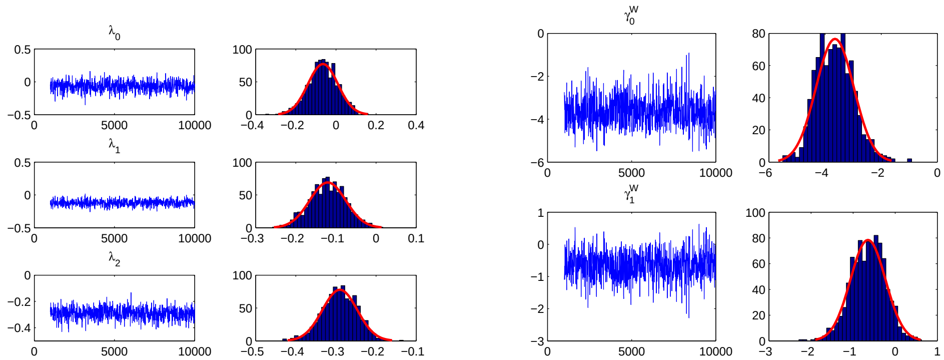

Figure 3.7: Chains and histograms for the parameter estimates of the mean and variance, cancer data

3.5.2 Example 2

In this example the data base has 312 registers with variables referring to the behavior of patients in a period from the time they responded to an ulcer treatment to the time the ulcer syptoms appear again or the closure of the study. The explained variable Y is the period of time just described. The covariates are $x_1 = 1, 2$, '1' if the symptoms of the ulcer appear again, '2' if not; $x_2 = 1, 2, 3, 4$ which is the time of the response treatment to the ulcer symptoms corresponding to 2, 4, 6 and 8 weeks; $x_3 = 1, 2$, '1' if the patient has quit smoking, '2' if not; $x_4 =$ grams of dayly consumption of alcohol; $x_5 = 0, \ldots, 9$ which is a graduation of coffee consumption from '0' very little to '9', a lot; $x_5 = 0, \ldots, 9$, a graduation of antiacids consumption from very little to a lot. To fit the data we tried a mixture of a normal and a weibull distribution modelling the mean, the variance and the incidence probability as follows

$$
\begin{aligned}
\mu_N(x_1,\ldots,x_6) &= \beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6 \qquad (3.24)\\
\mu_g(x_1,\ldots,x_6) &= \lambda_0+\lambda_1x_1+\lambda_2x_2+\lambda_3x_3+\lambda_4x_4+\lambda_5x_5+\lambda_6x_6\\
\sigma_N^2(x_1,\ldots,x_6) &= \exp\left(\gamma_0^N+\gamma_1^Nw_1+\gamma_2^Nw_2+\gamma_3^Nw_3+\gamma_4^Nw_4+\gamma_5^Nw_5+\gamma_6^Nw_6\right)\\
\sigma_g^2(x_1,\ldots,x_6) &= \exp\left(\gamma_0^w+\gamma_1^ww_1+\gamma_2^ww_2+\gamma_3^ww_3+\gamma_4^ww_4+\gamma_5^ww_5+\gamma_6^ww_6\right)\\
a_1(x_1) &= \frac{\exp(\tau_0+\tau_1x_1)}{1+\exp(\tau_0+\tau_1x_1)}
\end{aligned}
$$

where $\mathbf{w}=\mathbf{x}$. 7000 observations of the posterior distributions were generated, choosing every tenth observation with a burning period of 1000 observations. The results are given in table (3.4) and graphs and histograms of the chains are shown from figure (3.5.2) to figure(3.5.2). The estimation for $\boldsymbol{\tau}$ was $\hat{\boldsymbol{\tau}}=\left[\underset{(1.4721)}{5.0368},\underset{(1.4198)}{-5.8972}\right]$ with a 95% confidence interval for τ_1 rejecting the hypothesis, $\tau_1=0$. 95% confidence intervals for the parameters of the normal distribution rejected the individual hypothesis, $\beta_i=0$, for β_2, β_3 and β_4, all having negative signs, meaning that the time to feel ulcerous syntomatology again diminishes if the patient was slow to respond to the former ulcer treatment, or if he has not quit smoking, or as the consumption of alcohol increases. None of the parameters of the variance of the normal distribution rejected de null, $\gamma_i^N=0$. 95% confidence intervals for the parameters of the mean of the weibull distribution rejected the null, $\lambda_i=0$, for λ_1, λ_2, λ_3 and λ_4, all with negative signs, the meaning as for the parameters of the normal distribution. For the parameters of the variance of the weibull distribution, $\gamma_1^w>0$ and $\gamma_2^w<0$ were significantly different from zero. The only parameters which rejected the Jarque-Bera normality test were γ_2^w and γ_4^w.

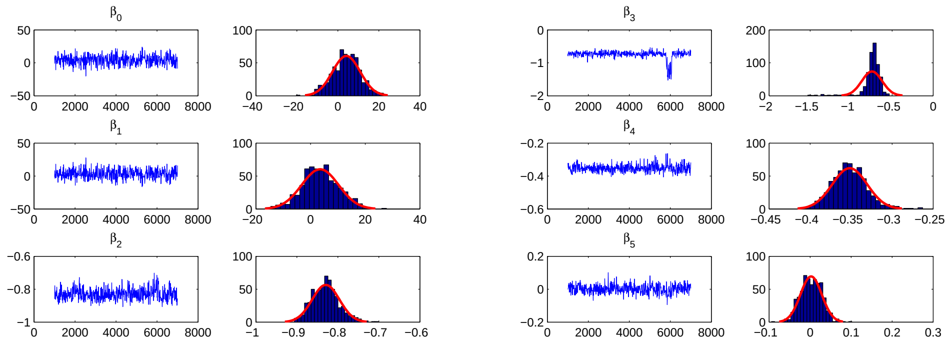

Figure 3.8: Chains and histograms for the parameter estimates of $\beta_0, \ldots, \beta_5$, example 2

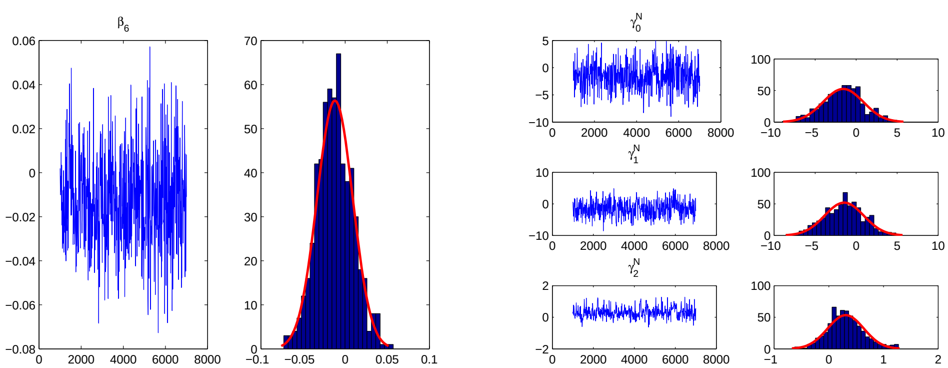

Figure 3.9: Chains and histograms for the parameter estimates of $\gamma_0^N, \ldots, \gamma_2^N$, example 2

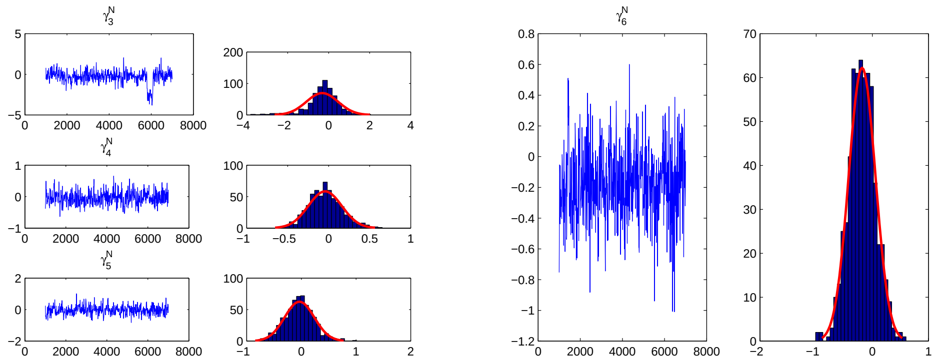

Figure 3.10: Chains and histograms for the parameter estimates of $\gamma_3^N, \ldots, \gamma_6^N$, example 2

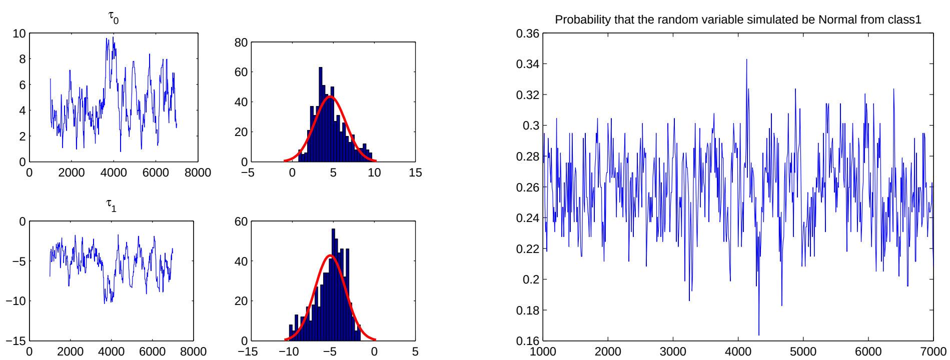

Figure 3.11: Chains and histograms for the parameter estimates of $\boldsymbol{\tau}$ and probability of being normal, example 2

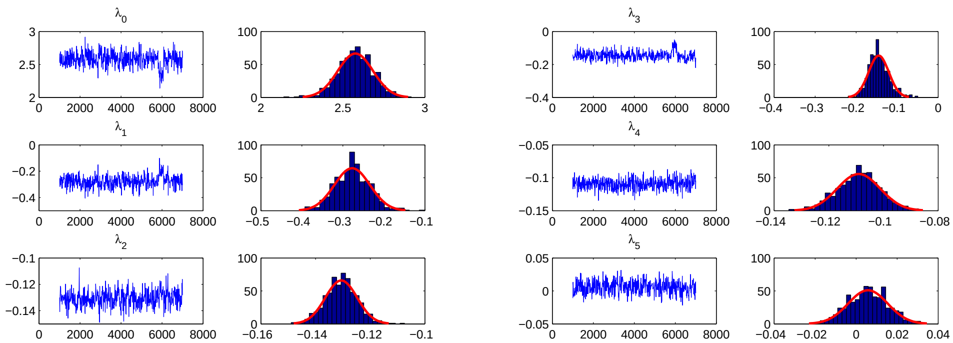

Figure 3.12: Chains and histograms for the parameter estimates of $\lambda_0, \ldots, \lambda_5$, example 2

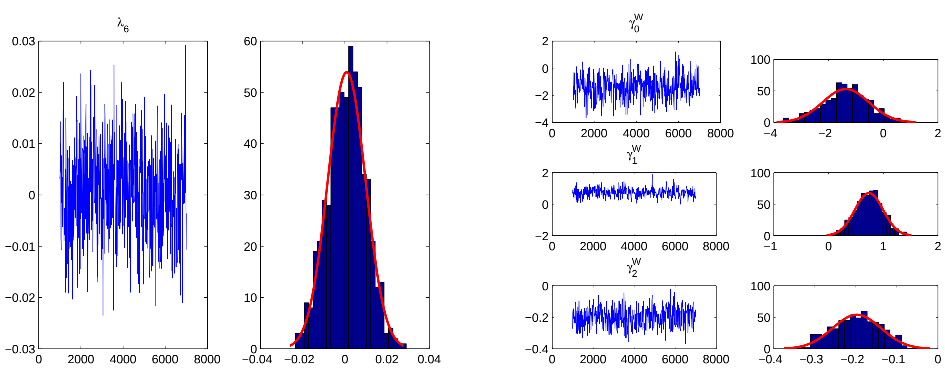

Figure 3.13: Chains and histograms for the parameter estimates of $\gamma_0^w, \ldots, \gamma_2^w$, example 2

$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$
4.0574 (6.9132)	3.5212 (6.8773)	−0.8340 (0.0368)	−0.7205 (0.0609)	−0.3496 (0.0246)	0.0086 (0.0217)	−0.0082 (0.0222)
$\hat{\gamma}_0^N$	$\hat{\gamma}_1^N$	$\hat{\gamma}_2^N$	$\hat{\gamma}_3^N$	$\hat{\gamma}_4^N$	$\hat{\gamma}_5^N$	$\hat{\gamma}_6^N$
−1.8809 (2.3773)	−1.7099 (2.0094)	0.2560 (0.3482)	−0.2153 (0.5223)	−0.0024 (0.1812)	−0.0209 (0.2913)	−0.1149 (0.2366)
$\hat{\lambda}_0$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{\lambda}_4$	$\hat{\lambda}_5$	$\hat{\lambda}_6$
2.5797 (0.1126)	−0.2769 (0.0412)	−0.1294 (0.0056)	−0.1572 (0.0219)	−0.1066 (0.0090)	0.0029 (0.0090)	0.0011 (0.0097)
$\hat{\gamma}_0^w$	$\hat{\gamma}_1^w$	$\hat{\gamma}_2^w$	$\hat{\gamma}_3^w$	$\hat{\gamma}_4^w$	$\hat{\gamma}_5^w$	$\hat{\gamma}_6^w$
−1.3136 (0.8122)	0.7009 (0.2469)	−0.1842 (0.0551)	−0.3961 (0.2256)	−0.1207 (0.0828)	0.1245 (0.1160)	0.0706 (0.0945)

Table 3.4: Estimation results from example 2

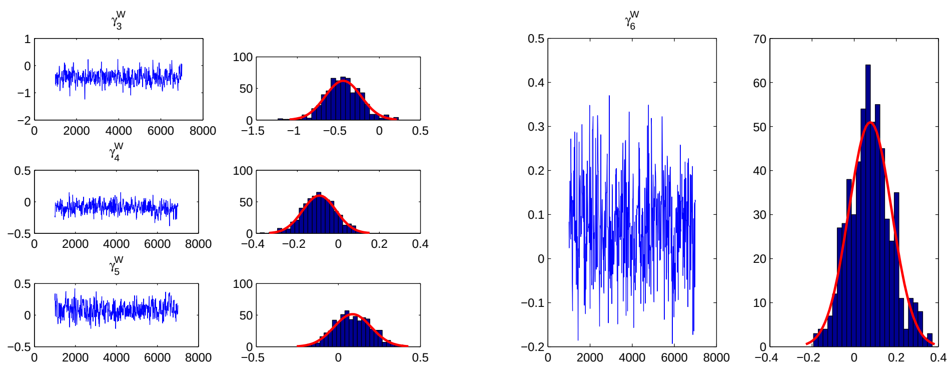

Figure 3.14: Chains and histograms for the parameter estimates of $\gamma_3^w, \ldots, \gamma_6^w$, example 2

3.6 Conclusions

The bayesian methodology to estimate the parameters in the mixture of normal and weibull distributions, modelling the mean and variance, showed good results as was seen in the simulations and the examples. The method can be extended to any of the extreme value distributions, or the generalized extreme value distribution, using the same techniques.

Bibliography

[1] Achcar, J. A.; Pereira, G. A., (1999a). Mixture model for type II survival in the presence of covarites, Computacional Statistics, 14: 223-250

[2] Achcar, J. A.; Pereira, G. A., (1999b) Use of exponential power distribution for mixture models in the presence of covariates, Journal of Applied Statistics, Vol. 26, N°6, 669-679

[3] Alexander C.(2001) Market Models, a Guide to Financial Data Analysis, first edition, John Wiley & Sons.

[4] Carrasco J.M.F.; Ortega E.M.M.; Paula G.A. (2008). Log-modified Weibull regression models with censored data: Sensitivity and residual analysis. Computational Statistics and Data Analysis 52 (2008) 4021-4039.

[5] Carroll, J. R., Roeder K. and Wasserman L. (1996). Flexible parametric measurement error models. Technical Reports N 648, Department of Statistics, Carnegie Mellon University.

[6] Casella G. ; Mengersen K. L; Robert C. P. ; D. M. Titterington D. M.(2002) Perfect Sampler for Mixtures of distributions Journal of the

Royal Statistics Society. Series B (Statistical Methodology), Vol. 64 N 4. pp 777-790

[7] Cepeda E.; Gamerman D (2001). Bayesian Modeling of Variance heterogeneity in Normal Regression Models. Brazilian Journal of Probability and Statistics, 14, pp. 207-221

[8] Cepeda E. and Gamerman D. (2005). Bayesian Methodology for modeling parameters in the two parameter exponential family. Estadstica, 57, 168 y 169, pp. 93 - 105.

[9] Chib, S. Greenberg, E (1995). the Metropolis-Hastings Algorithm. The American Statician 49(4), 327-335.

[10] Dey, D. K., Kuo, L. and Sahu, S. K. (1995). A Bayesian predictive approach to determining the number of components in a mixture distribution. Statistics and Computing, 5, 297-305.

[11] Diebolt J.; Robert C. P. (1994)Estimation of Finite Mixture Distributions through Bayesian Sampling. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 56, N 2. pp. 363-375

[12] Gamerman D.; Lopes H.F. (2006) Markov Chain Monte Carlo, Taylor & Francis.

[13] Gelfand, A. E.; Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association, 85, 398-409.

[14] Jiang R.; Murthy D.N.P. (1995). Modeling Failure-Data by Mixture of 2 Weibull Distributions: A Graphical Approach. IEEE transactions on reliability, vol. 44, No. 3 1995 September.

[15] Jiang R.; Murthy D.N.P. (1998). Mixture of Weibull distributions-Parametric characterization of failure rate function. Appl. Stochastic Models Data Anal. 14, 47-65 (1998).

[16] Mengersen, K. and Robert, C. P. (1996). Testing for mixtures: A Bayesian entropic approach. In Bayesian Statistics 5 (eds. J. O. Berger, J. M. Bernardo, A. P. Dawid, D. V. Lindley A. F. M. Smith. Oxford University Press: London.

[17] Peng F.; Robert J.A. ; Tanner M (1996) A. Bayesian Inference in Mixtures-of-Experts and Hierarchical Mixture-of-Expert Model With an Application to Speech Recognition. Journal of the American Statistical Association, Vol. 91, N435. pp. 953-960.

[18] Polymenis, A. and Titterington, D. M. (1998). On the determination of the number of components in a mixture. Statistics and Probability Letter, 38, 295-298.

[19] Richarson S.; Green P. J.(1997) On Bayesian Analysis of Mixtures with an Unknown Number of Components. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 59, N 4. pp. 731-792.

[20] Robert, C. P. (1996). Mixtures of distributions: inference and estimation. In Practical Markov Chain Monte Carlo, (eds. W. R. Gilks, S. Richardson, D. J. Spiegelhalter). London: Chapman and Hall.

[21] Roeder K. and Wasserman L. (1997). Practical Bayesian density estimation using mixtures of normals. Journal of the American Statistical Association, 92, 894-902.

[22] Sohn S.Y; Chang I.S.; Moon T.H (2006). Random effects Weibull regression model for occupational lifetime. European Journal of Operational Research 179 (2007) 124-131.

[23] Stephens, M. (1997). Bayesian methods for mixture of normal distributions. unpublished Ph.D. thesis, University of Oxford.

[24] Stephens, M. (1998). Bayesian analysis of mixtures with an unknown number of components - An alternative to reversible jump methods. 5

[25] Stephens, M. (2000). Dealing with label switching in mixture models. Journal of the Royal Statistical Society, B, 62, 795-809.

[26] Tanner M. A.; Wong W. H. (1987). The Calculation of Posterior Distributions by Data Augmentation. Journal of the American Statistical Association, Vol. 82, N398. pp. 528-540.

[27] Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). Statistical analysis of finite mixture distributions. Wiley: New York.

Chapter 4

MIXTURE MODELS APPLIED TO TAR MODELS AND SPATIAL STATISTICS

Summary: In this chapter the mixture of distributions is applied to TAR models and Spatial Statistics, using benchmark examples. *Key-words*:mixture models, variance heterogeneity, Bayesian methods, MCMC simulation.

4.1 Introduction

The purpose of this chapter is to take known examples from these two important fields in statistics, and use mixture models to fit them, giving explicit possible applications, based on the cooresponding field applications. So, there will not be any treatment on the theory in any of these fields.

4.2 Application to TAR models

TAR models have been used to describe the behavior of nonlinear time series with step autoregressive models, steps separated by estimable thresholds. This way the data are assigned to different regimes, that is, to different AR processes. These models have been introduced and developed by Tong (1978), Tong and Lim (1980), Tsay (1989), and the theme is extensively covered in Tong's book (1990). The usual estimation procedures are the maximum likelihood method and the conditional least square method (Tsay.1998). Bayesian estimaton procedures have also been used by Chen (1998), Chen (2005). Sangyeol et al, (1999) use sequential point estimation in a TAR(1) model. The proposal with mixture models is to explain the data with a mixture of normal distributions based on the assumption of normality of the error terms. There will not be an explicit estimation of the threshold, but the probability of belonging to any of the Normal distributions in the mixture will depend on some lags of the time series variable. In order to be able to compare with other models, the applications are the Canadian lynx data set and the sunspots data set, well known examples in the time series literature. A mixture model called MixAR has already been proposed by Zeevi A.J. et al (1998) using maximum likelihood methods for the estimation procedure.

4.2.1 TAR models

The underlying idea in TAR models, is the use of piecewise linear approximations to analyze a non linear time series.

A TAR model has the form

$$y_t = a_0^{(j)} + \sum_{i=1}^{m_j} a_i^{(j)} y_{t-i} + \sum_{i=0}^{m'_j} b_i^{(j)} x_{t-i} + \epsilon_t^{(j)}$$

$$\text{if} \quad r_{j-1} < z_{t-d} \le r_j$$

where $j = 1, ..., l$, $-\infty = r_0 < r_1 < ... < r_{l-1} < r_l = \infty$, $\{y_t\}$ is the output time sries, $\{x_t\}_{t-m'_j}^{t}$, $j = 1, ..., l$, is a set of exogenous variables, z_t is the threshold variable, whose value indicates which linear equation is in use. z_t is supposed known, stationary and has a continuous distribution. $\left\{\epsilon_t^{(j)}\right\}$ is a sequence of non correlated random variables with mean zero, finite variances, and independent from $\{x_t\}_{t-m'_j}^{t}$. The parameters to be estimated are $a_i^{(j)}$, $i = 0, ..., m_j$, $j = 1, ..., l$, $b_i^{(j)}$, $i = 0, ..., m'_j$, $j = 1, ..., l$, r_j, $j = 1, ..., l-1$, d and l.

A particular case of the TAR models is when the threshold variable is the time series variable, a self excited TAR model, called SETAR. This is the model used to fit the sunspot and Lynx data sets. The mixtures to be applied, are based on these models.

4.3 The mixture model

Let $\{Y_t\}$ be the stochastic process. The density function of Y_t conditonal on $\{y_1, \ldots, y_{t-1}\}$ is given by

$$f(y_t \,|\boldsymbol{\theta}, y_{t-1}, \ldots, y_1) = \sum_{j=1}^{k} a_j f_j(y_t; \theta_j, y_{t-1}, \ldots, y_1). \tag{4.1}$$

We specifically assume the model where the mixture components and the weights are given by

$$f_j(y_t\,|\boldsymbol{\theta}_j, y_{t-1},\ldots,y_1) = \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{1}{2}\left(\frac{y_t-(\mathbf{L_j(y_t;}\boldsymbol{\beta}_\mathbf{j}))}{\sigma_j}\right)^2} \tag{4.2}$$

$$a_j = \frac{e^{L_j(y_t;\boldsymbol{\tau}_j)}}{1+\sum_{l=1}^{l=k-1} e^{\mathbf{L_l(y_t;}\boldsymbol{\tau}_\mathbf{l})}} \tag{4.3}$$

where $j = 1, ..., k$ for (4.2) ; $j = 1, \ldots, k-1$ for (4.3), $a_k = 1 - \sum_{j=1}^{k-1} a_j = \frac{1}{1+\sum_{l=1}^{l=k-1} e^{\mathbf{L_l(y_t;}\boldsymbol{\tau}_\mathbf{l})}}$, $L_j(y_t; \boldsymbol{\beta}_j) = \beta_0 + \beta_1 y_{t-1} + \ldots \beta_{m_j} y_{t-m_j}$ and $L_j(y_t; \boldsymbol{\tau}_j) = \tau_0 + \tau_1 y_{t-1} + \ldots \tau_{n_j} y_{t-n_j}$. That is, with probability $a_j(L_j(y_t; \boldsymbol{\tau}_j))$, Y_t follows an AR(m_j) process

$$y_t = \beta_0 + \beta_1 y_{t-1} + \ldots \beta_{m_j} y_{t-m_j} + e_j$$

where $e_j \sim N(0, \sigma_j^2)$.

This mixture of normal distributions is a special case of the mixture of distributions from the biparametric exponential family in chapter two, so, the same bayesian methodology is employed to estimate the parameters, with the lagged variables as covariates and constant variances $\sigma_j^2 = e^{\gamma_j}$.

4.3.1 Canadian Lynx data set

This set consists of the number of lynx trapped in the Mackenzie River district of North West Canada from 1821 to 1934. The results of the Bayesian procedure for this example will be compared with those of Tong (1990), thus the $\log_{10}$ transformation of the data will be used as well as the same AR models. A first linear time series model suggested by Moran (1953) was the

AR(2) model whose estimation by standard parametric procedures was

$$Y_t = 1.05 + 1.41Y_{t-1} - 0.77Y_{t-2} + \epsilon_t$$

where $\epsilon_t \sim$IID$(0, 0.04591)$.

Using the bayesian estimation procedure for mixture models (with just one distribution), described in previous chapters, the results are

$$\underset{(0.1266)}{1.0587} + \underset{(0.0663)}{1.3804}Y_{t-1} - \underset{(0.0666)}{0.7442}Y_{t-2} + \epsilon_t$$

where $\epsilon_t \sim$IID$(0, e^{-2.9188})$=IID$(0, 0.054)$. 95% credible intervals for each of the parameters contain the values estimated by Moran (1953). The BIC value for this model is BIC=24.87. Figure (4.3.1) shows the corresponding chains and histograms.

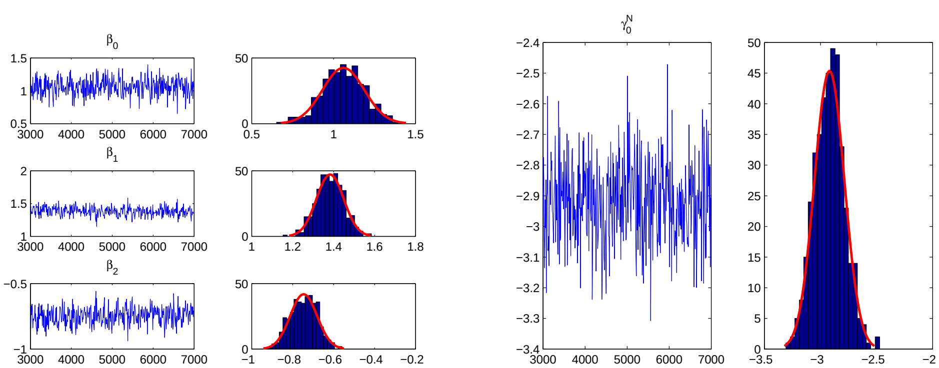

Figure 4.1: Chains and histograms for the parameter estimates of the AR(2) model of the Canadian Lynx data set

A well fitted model proposed by Tong (1990) is the SETAR(2;7,2) given by

$$Y_t = \begin{cases} \underset{(0.275)}{0.546} + \underset{(0.094)}{1.032}Y_{t-1} - \underset{(0.156)}{0.173}Y_{t-2} + \underset{(0.149)}{0.171}Y_{t-3} - \underset{(0.153)}{0.431}Y_{t-4} & \\ + \underset{(0.170)}{0.332}Y_{t-5} - \underset{(0.167)}{0.284}Y_{t-6} + \underset{(0.101)}{0.210}Y_{t-7} + \epsilon_t^{(1)} & \text{if } Y_{t-2} \leq \underset{[2.926,3.123]}{3.116} \\ \underset{(0.655)}{2.632} + \underset{(0.102)}{1.492}Y_{t-1} - \underset{(0.195)}{1.324}Y_{t-2} + \epsilon_t^{(2)} & \text{if } Y_{t-2} > 3.116 \end{cases} \tag{4.4}$$

with $\text{var}(\epsilon_t^{(1)}) = 0.0259$ and $\text{var}(\epsilon_t^{(2)}) = 0.0505$, BIC=-298.4.

The results of the Bayesian estimation of a mixture of two normal distributions, one with the first seven lags as covariates, the second with the first two lags as covariates and no covariates explaining the incidence probability, $a_1 = \frac{e^\tau}{1+e^\tau}$, is

$$Y_t = \begin{cases} \underset{(0.4409)}{1.3313} + \underset{(0.1320)}{1.2718}Y_{t-1} - \underset{(0.2460)}{0.8753}Y_{t-2} + \underset{(0.3081)}{0.4964}Y_{t-3} - \underset{(0.1985)}{0.5634}Y_{t-4} & \\ + \underset{(0.2190)}{0.094}Y_{t-5} - \underset{(0.2190)}{0.2670}Y_{t-6} + \underset{(0.1605)}{0.3623}Y_{t-7} + \epsilon_t^{(1)} & \gamma^{N_1} = \underset{(0.3907)}{-3.6635} \\ \underset{(0.1479)}{0.8392} + \underset{(0.0914)}{1.0841}Y_{t-1} - \underset{(0.1032)}{0.3321}Y_{t-2} + \epsilon_t^{(2)} & \gamma^{N_2} = \underset{(0.4326)}{-4.0921} \end{cases} \tag{4.5}$$

where Y_t is modelled by the first equation with probability a_1 and by the second equation with probability $1 - a_1$, with $\hat{\sigma_1^2} = e^{-3.6635} = 0.026$, $\hat{\sigma_2^2} = e^{-4.0921} = 0.017$, $\hat{\tau} = \underset{(0.5497)}{0.1301}$, so, $a_1 = \frac{e^{0.1301}}{1+e^{0.1301}} = 0.53$ which means that approximately 50% of the obsevations belong to the first normal distribution. The BIC value was BIC=85.0646. As can be seen the signs in all the coefficients coincide with those of Tong in (4.4) but the BIC value is much greater than that obtained by Tong.

In search of a better estimation a second model is estimated, this one differing from the one above just in the incidence probability now being explained by the first three lags, $a_1 = \frac{e^{\tau_0+\tau_1 Y_{t-1}+\tau_2 Y_{t-2}+\tau_3 Y_{t-3}}}{1+e^{\tau_0+\tau_1 Y_{t-1}+\tau_2 Y_{t-2}+\tau_3 Y_{t-3}}}$, the results are

$$Y_t = \begin{cases} \underset{(0.5885)}{1.0318} + \underset{(0.1538)}{1.2934} Y_{t-1} - \underset{(0.3085)}{1.0800} Y_{t-2} + \underset{(0.3763)}{0.8436} Y_{t-3} - \underset{(0.2931)}{0.6444} Y_{t-4} \\ + \underset{(0.3192)}{0.0970} Y_{t-5} - \underset{(0.2969)}{0.2074} Y_{t-6} + \underset{(0.2078)}{0.3272} Y_{t-7} + \epsilon_t^{(1)} \quad \gamma^{N_1} = \underset{(0.3315)}{-2.9719} \\ \underset{(0.1712)}{0.8392} + \underset{(0.1448)}{1.0171} Y_{t-1} - \underset{(0.1483)}{0.2172} Y_{t-2} + \epsilon_t^{(2)} \quad \gamma^{N_2} = \underset{(0.2922)}{-4.0663} \end{cases} \tag{4.6}$$

where Y_t is modelled by the first equation with probability a_1 and by the second equation with probability $1 - a_1$, with $\hat{\sigma_1^2} = \underset{(0.0175)}{0.0480}$, $\hat{\sigma_2^2} = \underset{(0.0068)}{0.0189}$, $\hat{\tau_0} = \underset{(2.6106)}{-7.1536}$, $\hat{\tau_1} = \underset{(1.8934)}{-3.3490}$, $\hat{\tau_2} = \underset{(1.6663)}{1.8326}$, $\hat{\tau_3} = \underset{(2.4650)}{4.1698}$. 95% credible intervals reject the individual hypothesis $\tau_1 = 0$ and $\tau_3 = 0$, so the lags 1 and 3 explain the change from one distribution to the other. The BIC value was BIC=-216.8488, which improved considerably in spite of having 4 more parameters. 15000 observations were generated and the estimations were calculated with the last 5000, figures (4.3.1) to (4.3.1) show some of the chains and histograms.

4.3.2 Sunspot numbers

This data set consits of the annual sunspot numbers from 1700 to 1988. The models developed here are based on SETAR models suggested in Tong's book (1990) and are applied to a square root transformation of the original data. The first model suggested is a reparametrization (1983) of a model originally

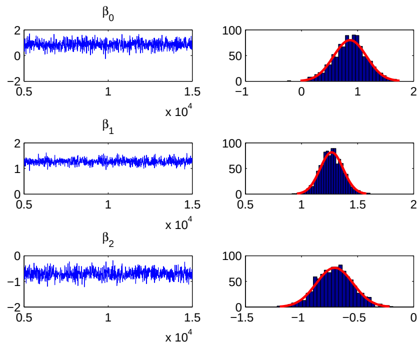

Figure 4.2: Chains and histograms for β_0, β_1, β_2 of the first normal distribution, Lynx data set

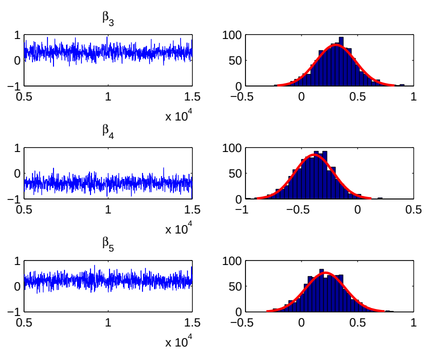

Figure 4.3: Chains and histograms for β_3, β_4, β_5 of the first normal distribution, Lynx data set

proposed by Ghaddar and Tong (1980):

$$Y_t = \begin{cases} 1.89 + 0.86Y_{t-1} + 0.08Y_{t-2} - 0.32Y_{t-3} + 0.16Y_{t-4} - 0.21Y_{t-5} & \\ -\,0.00Y_{t-6} + 0.19Y_{t-7} - 0.28Y_{t-8} + 0.2Y_{t-9} + 0.1Y_{t-10} + \epsilon_t^{(1)} & \text{if } Y_{t-8} \leq 11.93 \\ 4.53 + 1.41Y_{t-1} - 0.78Y_{t-2} + \epsilon_t^{(2)} \quad \text{if } Y_{t-8} > 11.93 & \end{cases} \tag{4.7}$$

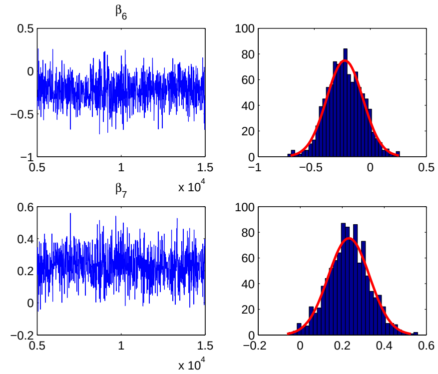

Figure 4.4: Chains and histograms for β_6, β_7 of the first normal distribution, Lynx data set

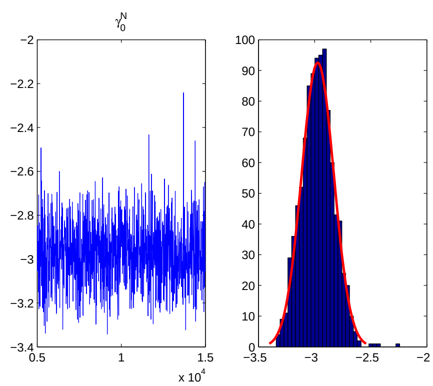

Figure 4.5: Chain and histogram for γ of the first normal distribution, Lynx data set

$\text{var}\epsilon_t^{(1)} = 1.946$, $\text{var}\epsilon_t^{(2)} = 6.302$

A Bayesian estimation of a mixture of two normal distributions, the first one with the first 10 lags as covariates, the second one with the first two lags

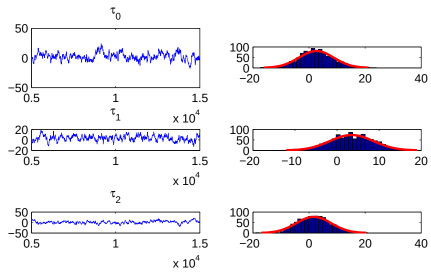

Figure 4.6: Chains and histograms for τ_0, τ_1, τ_2, incidence probability, Lynx data set

as covariates and the incidence probability explained by the eight lag is:

$$Y_t = \begin{cases} \underset{(0.27)}{1.1658} + \underset{(0.0774)}{0.8813} Y_{t-1} - \underset{(0.1066)}{0.0084} Y_{t-2} - \underset{(0.1020)}{0.2792} Y_{t-3} + \underset{(0.0924)}{0.2022} Y_{t-4} \\ - \underset{(0.0915)}{0.2252} Y_{t-5} - \underset{(0.0907)}{0.0453} Y_{t-6} + \underset{(0.0892)}{0.2272} Y_{t-7} - \underset{(0.0891)}{0.3060} Y_{t-8} \\ + \underset{(0.0913)}{0.1592} Y_{t-9} + \underset{(0.0623)}{0.1264} Y_{t-10} + \epsilon_t^{(1)} \quad \gamma^{N_1} = \underset{(0.1331)}{-0.7351} \\ \underset{(0.3943)}{2.9990} + \underset{(0.0719)}{1.3954} Y_{t-1} - \underset{(0.0949)}{0.8028} Y_{t-2} + \epsilon_t^{(2)} \quad \gamma^{N_2} = \underset{(0.1691)}{0.5070} \end{cases} \tag{4.8}$$

where Y_t is modelled by the first equation with probability a_1 and by the second equation with probability $1 - a_1$, $\text{var}\epsilon_t^{(1)} = \underset{(0.0658)}{0.4803}$, $\text{var}\epsilon_t^{(2)} = \underset{(0.2927)}{1.6761}$. The incidence probability is $a_1 = \frac{e^{\tau_0 + \tau_8 Y_{t-8}}}{1+e^{\tau_0 + \tau_8 Y_{t-8}}}$, where $\tau_0 = \underset{(3.0756)}{11.4191}$, $\tau_8 = \underset{(0.4403)}{-1.5242}$, the BIC value is BIC=-318.2172. Chains and histograms are shown in figures (4.3.2) to (4.3.2). As can be seen, all but the sign of Y_{t-2}, which is not significantly different from zero, coincide with those of Tong.

The Bayesian methodology was also applied to mixture models corresponding to AR(9) and SETAR(2;3,11) suggested in Tong's book (1990), with the incidence probability of the second model explained by the third

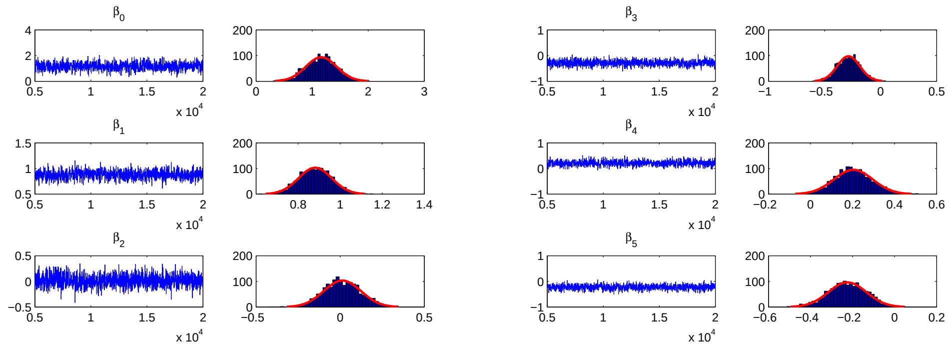

Figure 4.7: Chains and histograms for $\beta_0 \ldots \beta_5$, of the first normal distribution, Sunspot data set

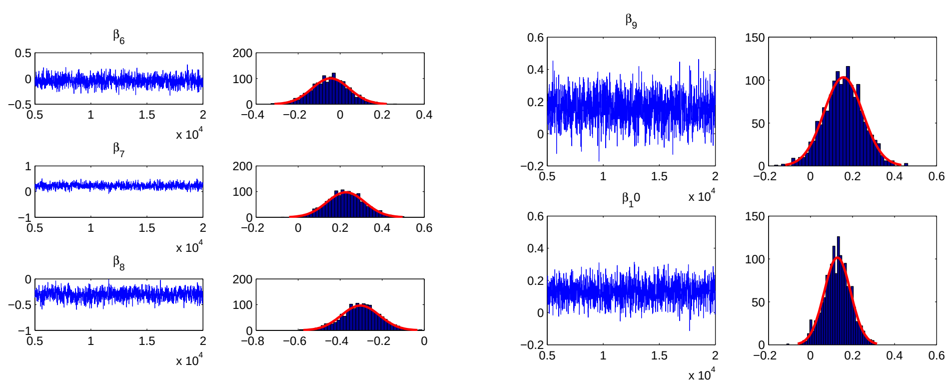

Figure 4.8: Chains and histograms for $\beta_6 \ldots \beta_{10}$, of the first normal distribution, Sunspot data set

lag. Table (4.3.2) below shows the results of predictions from 1981 to 1988 for the three models. Following Tong's notation, the models are denoted mix(2;10,2), AR(9) and mix(2;3,11). 95% confidence bands of prediction are shown in figure (4.3.2). The dots in the figures represent the true observation.

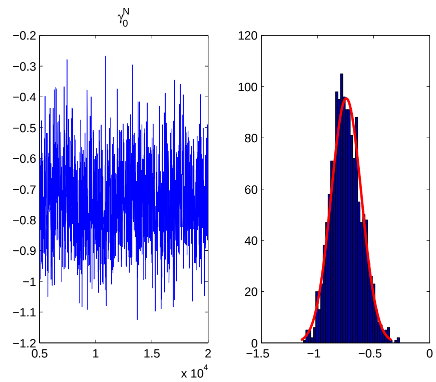

Figure 4.9: Chains and histograms for γ, of the first normal distribution, Sunspot data set

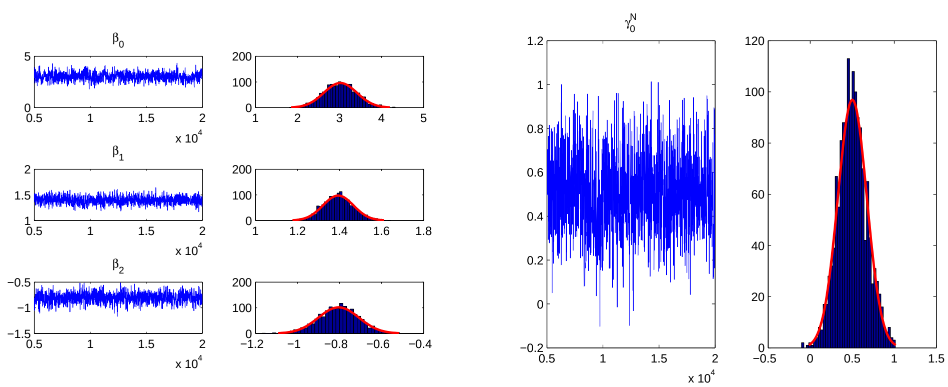

Figure 4.10: Chains and histograms for $\beta_0 \dots \beta_2$, and γ of the second normal distribution, Sunspot data set

4.4 Application to Spatial Statistics

In this section the mixture of two distributions with joint modelling of the mean and the variance, using bayesian methodology to estimate the parameters, is employed to fit a growth model for the regions of the European Union classified in NUTS level 2. The data base with the variables Gross

Year	True value	mix(2;10,2)	AR(9)	mix(2;3,11)
1980	154.7	155.462 (7.95246)	151.707 (6.30474)	154.057 (6.70153)
1981	140.5	128.353 (9.34266)	123.505 (9.56811)	140.330 (7.26869)
1982	115.9	89.314 (8.63579)	83.884 (10.07161)	92.735 (7.12004)
1983	66.6	56.906 (6.24949)	54.354 (8.35017)	65.074 (6.31592)
1984	45.9	30.188 (4.09290)	29.540 (6.19845)	31.248 (4.78535)
1985	17.9	16.241 (2.83858)	17.884 (4.78756)	20.542 (3.59928)
1986	13.4	16.989 (3.01240)	14.161 (4.12074)	9.347 (2.29460)
1987	29.2	30.381 (5.50229)	21.840 (5.10707)	9.817 (2.39519)
BIC		−318.2172	682.2574	436.6435
MSE		151.60293	224.39347	144.14611

Table 4.1: Table of predictions for the sunspot data

Domestic Product at current market prices, and growth rate of regional Gross Value Added (GVA) at basic prices (percentage change on previous year), was obtained from the Eurostat data base. The data base in the model to be fitted is cross-sectional, taking base year 2006 for the GDP and 2007 for the growth rate, considering the regions in the NUTS 2 classification which had no missing values for these years, resulting in 190 regions which ecompassed all the 27 countries in the European Union. The aim of this section is to show an application of the mixture of normal and gamma distributions to a known spatial econometric example as it is the β-convergence, from the statistical point of view, not including detailed economic analyses, which are

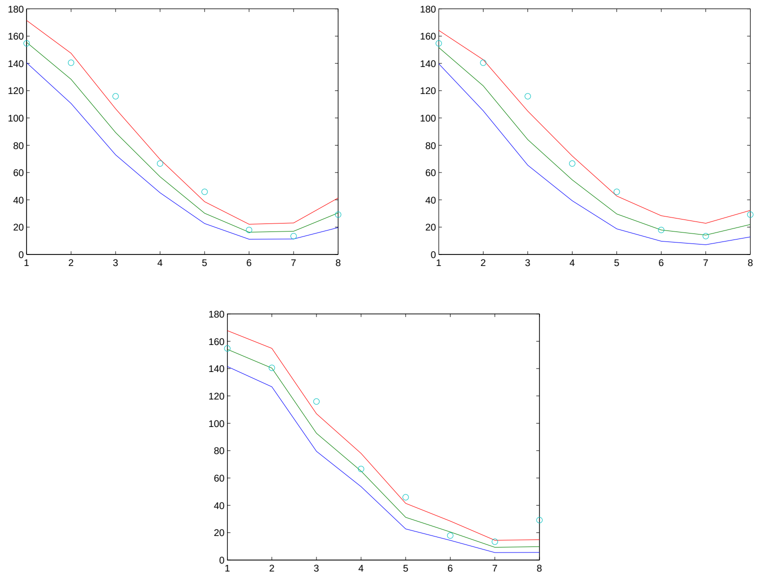

Figure 4.11: 95% confidence bands for the predictions in models mix(2;10,2) (left), AR(9) (right) and mix(2;3,11) (bottom)

worth analizing in a further study. The spatial weight matrix W to be used is very simple, $w_{ii} = 0$, $w_{ij} = 1$ if regions i and j belong to the same country and $w_{ij} = 0$ if they belong to different countries. The matrix is normalized so that the elements of each row sum to one.

4.4.1 The model

The names of the variables are the same as used in the literature of the β-convergence model, that is, g =growth rate, $\ln y$ =log(GDP). Four models are fitted, using the BIC value to compare them.

1. In this model neither the spatial correlation nor the heteroscedasticity are taken into account. g follows a normal distribution, $N(\mu_N, \sigma_N^2)$,

with probability a_1, and a gamma distribution, $\Gamma(\alpha, \delta)$, with probability $1 - a_1$, where

$$\begin{aligned}
\mu_N &= \beta_0 + \beta_1 \ln y, \\
\sigma_N^2 &= \exp\left(\gamma_0^N\right), \\
\mu_g &= \alpha\delta = \lambda_0 + \lambda_1 \ln y, \\
\sigma_g^2 &= \alpha\delta^2 = \exp\left(\gamma_0^g\right), \\
a_1 &= \frac{e^{\tau_0}}{1 + e^{\tau_0}}
\end{aligned} \tag{4.9}$$

2. In this model spatial coorelation as heteroscedasticity are considered, but with one normal distribution in the mixture. g follows a normal distribution, $N(\mu_N, \sigma_N^2)$, with probability $a_1 = 1$, where

$$\begin{aligned}
\mu_N &= \beta_0 + \beta_1 \ln y + \beta_2 W \ln y + \beta_3 W g, \\
\sigma_N^2 &= \exp\left(\gamma_0^N + \gamma_1^N \ln y + \gamma_2^N W \ln y + \gamma_3^N W g\right)
\end{aligned} \tag{4.10}$$

3. The third model is homoscedastic with spatial correlation. g follows a normal distribution, $N(\mu_N, \sigma_N^2)$, with probability a_1, and a gamma distribution, $\Gamma(\alpha, \delta)$, with probability $1 - a_1$, where

$$\begin{aligned}
\mu_N &= \beta_0 + \beta_1 \ln y + \beta_2 W \ln y + \beta_3 W g, \\
\sigma_N^2 &= \exp\left(\gamma_0^N\right), \\
\mu_g &= \lambda_0 + \lambda_1 \ln y + \lambda_2 W \ln y + \lambda_3 W g, \\
\sigma_g^2 &= \exp\left(\gamma_0^g\right), \\
a_1 &= \frac{e^{\tau_0 + \tau_1 \ln y}}{1 + e^{\tau_0 + \tau_1 \ln y}}
\end{aligned} \tag{4.11}$$

4. The fourth model is heteroscedastic with spatial correlation and g follows a normal distribution, $N(\mu_N, \sigma_N^2)$, with probability a_1, and a gamma distribution, $\Gamma(\alpha, \delta)$, with probability $1 - a_1$, where

$$\begin{aligned}
\mu_N &= \beta_0 + \beta_1 \ln y + \beta_2 W \ln y + \beta_3 W g, \qquad (4.12) \\
\sigma_N^2 &= \exp\left(\gamma_0^N + \gamma_1^N \ln y + \gamma_2^N W \ln y + \gamma_3^N W g\right), \\
\mu_g &= \lambda_0 + \lambda_1 \ln y + \lambda_2 W \ln y + \lambda_3 W g, \\
\sigma_g^2 &= \exp\left(\gamma_0^g + \gamma_1^g \ln y + \gamma_2^g W \ln y + \gamma_3^g W g\right), \\
a_1 &= \frac{e^{\tau_0 + \tau_2^g W \ln y + \tau_3^g W g}}{1 + e^{\tau_0 + \tau_2^g W \ln y + \tau_3^g W g}}
\end{aligned}$$

4.4.2 Estimation results

15000 observations from the posterior distributions of the parameters were generated, choosing every tenth observation to form the chains. The results are shown in table (4.4.2).

All the parameters in the first model are highly significant, using 95% credible intervals. $\hat{\sigma}_N^2 = 1.9662(0.2803)$, $\hat{\sigma}_g^2 = 4.8567(1.4080)$ and $\hat{\boldsymbol{\tau}} = \left[\underset{(2.2903)}{-22.1200}, \underset{(0.2395)}{2.4546}\right]'$. That is, the variability in the rate of growth of the countries which belong to the first distribution with probability a_1, is lower than those which belong to the second distribution with probability $1 - a_1$, and the probability to belong to the normal distribution increases when the GDP increases.

In the second model, the only parameter not significantly different from zero is γ_1^N, any way, showing heteroscedasticity.

In the third model the BIC value improved significantly; the only parameters not significantly different from zero are β_1 and λ_2. $\hat{\sigma}_N^2 = 7.8135(1.4336)$,

$\hat{\sigma}_g^2 = 0.5726(0.0886)$ and $\hat{\boldsymbol{\tau}} = \left[\underset{(1.0872)}{1.2996}, \underset{(0.6944)}{3.4300}, \underset{(0.6977)}{-4.3450}, \underset{(0.2023)}{1.2511}\right]'$.

The only parameter not significantly different from zero in the fourth model is γ_1^g. It has a BIC value lower than the other models, and shows heteroscedasticity and spatial correlation. $\hat{\boldsymbol{\tau}} = \left[\underset{(2.0697)}{-5.9417}, \underset{(0.6410)}{2.3888}, \underset{(1.9763)}{-5.0869}\right]$.

In the theory of β-convergence, a negative value of the coefficient of $\ln y$ means that regions with lower GDP in the base year have greater rates of growth than richer regions, fact which eventually will reduce the gap between poor and rich regions, implying a convergence to a steady state of the economy. Many authors have obtained this negative coefficient with a period of time longer than five years. Among the authors who have worked the β-convergence model are Moreno et al, (2000), Baumont C. et al (2001) and Dall'erba S. (2005), using classical spatial econometric techniques; Crespo J. et al, (2010) employ Bayesian methodology to estimate the parameters. In all the models presented here, the values of β_1 and λ_1, the coefficients of $\ln y$, are positive, maybe because of the short period of time, one year, or because, by this time, the European Union is more consolidated. A better interpretation of this result requires further studies of models with longer and different periods of time as well as with different weight matrices. As has been said before, the aim is to show the versatility of the mixture models.

Chains from model (4) are shown from figure (4.4.2) to figure (4.4.2).

	Model(1)	Model(2)	Model(3)	Model(4)
$\hat{\beta}_0$	-7.0714 (2.8169)	-1.3400 (2.1139)	2.8408 (3.9787)	-6.6749 (2.6465)
$\hat{\beta}_1$	0.9558 (0.2806)	0.5983 (0.2350)	0.2361 (0.4083)	1.1262 (0.3092)
$\hat{\beta}_2$		-0.4810 (0.0874)	-0.6382 (0.1287)	-0.4568 (0.1783)
$\hat{\beta}_3$		1.0496 (0.0764)	1.1193 (0.1725)	0.9470 (0.2047)
$\hat{\gamma}_0^N$	0.6771 (0.1314)	4.2930 (1.2712)	2.0366 (0.1824)	1.7349 (2.3388)
$\hat{\gamma}_1^N$		-0.1619 (0.1522)		0.9536 (0.3046)
$\hat{\gamma}_2^N$		-0.3342 (0.0599)		-1.3746 (0.1728)
$\hat{\gamma}_3^N$		0.2882 (0.0589)		1.0946 (0.1780)
$\hat{\lambda}_0$	-0.9538 (1.0789)		-3.6585 (0.7019)	0.1303 (0.5966)
$\hat{\lambda}_1$	0.3270 (0.1213)		0.3719 (0.0823)	0.1394 (0.0599)
$\hat{\lambda}_2$			-0.0155 (0.0826)	-0.0966 (0.0213)
$\hat{\lambda}_3$			0.3781 (0.0245)	0.2121 (0.0349)
$\hat{\gamma}_0^g$	1.5361 (0.2626)		-0.5861 (0.1534)	4.6164 (1.9262)
$\hat{\gamma}_1^g$				-0.2529 (0.2156)
$\hat{\gamma}_2^g$				-0.6108 (0.1336)
$\hat{\gamma}_3^g$				0.8165 (0.2053)
BIC	167.9013	236.5452	-623.2270	-741.3387

Table 4.2: Results from the spatial growth model

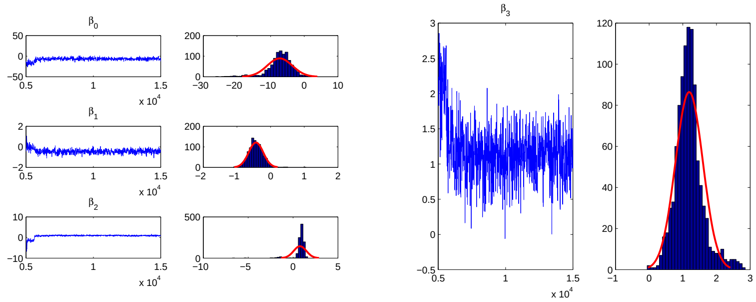

Figure 4.12: Chains and histograms for $\hat{\beta}_0 - \hat{\beta}_3$, spatial growth model

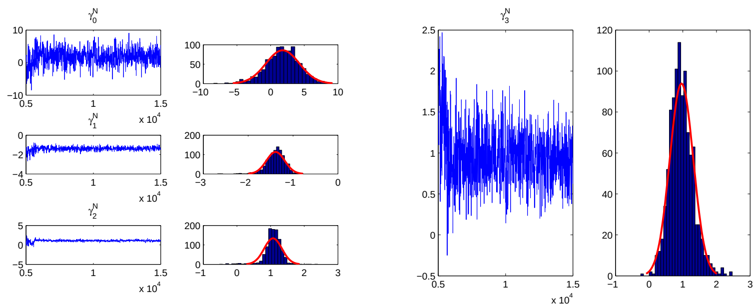

Figure 4.13: Chains and histograms for $\hat{\gamma_0^N} - \hat{\gamma_3^N}$, spatial growth model

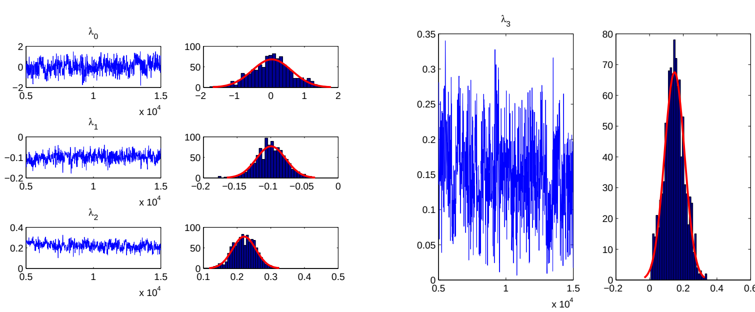

Figure 4.14: Chains and histograms for $\hat{\lambda}_0 - \hat{\lambda}_3$, spatial growth model

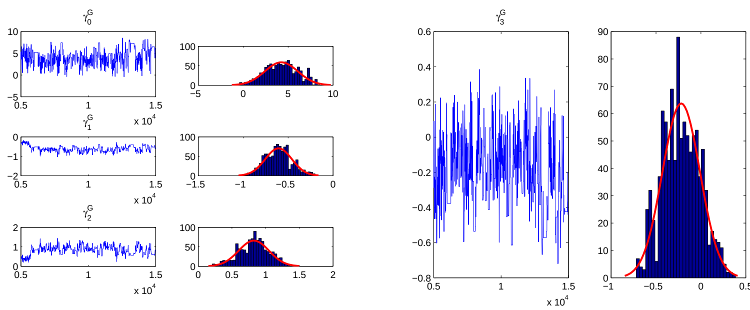

Figure 4.15: Chains and histograms for $\hat{\gamma}_0^g - \hat{\gamma}_3^g$, spatial growth model

Bibliography

[1] Alexander C.(2001) Market Models, a Guide to Financial Data Analysis, first edition, John Wiley & Sons.

[2] Baumont Catherine, Ertur Cem and Le Gallo Julie (2001). The European Regional Convergence Process, 1980-1995: do Spatial Regimes and Spatial Dependence matter? University of Burgundy, LATEC UMR-CNRS 5118, Pôle d'Economie et de Gestion.

[3] Chen Cathy W.S. & So Mike K.P.(2006) On a threshold heteroscedastic model. International Journal of Forecasting 22, 73-89. Published by Elsevier B.V.

[4] Crespo Cuaresma Jesús and Feldkircher Martin (2010). Spatial Filtering, Model Uncertainty and the speed of Income Convergence in Europe. OESTERREICHISCHE NATIONALBANK, working paper 160.

[5] Dall'erba Sandy (2003). Distribution of regional income and regional funds in Europe 1989-1999: An exploratory spatial data analysis. The Annals of Regional Science (2005) 39:121-148.

[6] Moran, P.A.P. (1953). The statistical analysis of the Canadian lynx cycle, I: structure and predicition. Aust. J. Zool, 1, 163-73.

[7] Moreno R. and Vayá E. Técnicas econométricas para el tratamiento de datos espaciales: La econometría espacial. Edicions de la Universitat de Barcelona (2000).

[8] Sangyeol L. & Sriram T.N.(1999) Sequential point estimation of parameters in a threshold AR(1) model. Stochastic Processes and their Applications 84 343-355. Elsevier Science B.V.

[9] Tsay R.S. (1998). Testing and Modeling Multivariate Threshold Models. Journal of the American Statistical Association, September 1998, Vol. 93, N443.

[10] Tong H. (1995). Non-linear Time Series A Dynamical System Approach. Oxford Science Publications.

[11] Zeevi A.J.,Meir R., Adler R.J. (1998). Non-Linear Models for Time Series Using Mixtures of Autoregressive Models

Conclusion

Bayesian methodology to estimate the parameters in finite mixture of distributions from the exponential family and finite mixture of normal and weibull distributions, in particular, the choice of the kernel distribution in the Metropolis-Hastings algorithm, showed good results in the convergence of the chains of the observations of the posterior distributions of the parameters involved for well separated mixtures. In cases of ill-separated components, the normal component tends to take more observations than initially assigned to this distribution in the simulation. Thus, when the sample size is small, the method would wrongly detect a single normal distribution, however for a large sample size the method separates the distributions and can estimate the parameters correctly.

The joint modelling of the mean and the variance proved to be useful for the treatment of heteroscedasticity, as was seen in many of the examples and the spatial statistics application. It was also seen that a finite mixture is, in many instances, a better model than a unimodal distribution.

As a theoretic selection criterion, the BIC was used throughout. Since it is widely used, it was possible to compare the performance of the proposed mixture models with classical models in some benchmark examples as it is the sunspot data set. A further study could establish differences among selection criteria for these mixture models, varying for example sample size, number of regressors, etc.. This comparisson could be between BIC and DIC, being DIC a modification of BIC, widely used in Bayesian theory nowadays, behaving asymptotically as BIC.

There was a general treatment in the choice of the kernel in the Metropolis-Hastings algorithm, so the bayesian methodology employed can be easily generalized to finite mixtures of distibutions belonging to two parametric families, which could have any number of parameters. This generalization also includes any functional modelling of the mean and the variance, as for example in ARCH and GARCH models.

It is also easy to generalize the algorithms to deal with any finite number of families of distributions in a finite mixture.